UMA ABORDAGEM BASEADA EM FLUXO DE FILTROS PARA O RECONHECIMENTO DE ENTIDADES EM MENSAGENS DO TWITTER

DIEGO MARINHO DE OLIVEIRA

UMA ABORDAGEM BASEADA EM FLUXO DE FILTROS PARA O RECONHECIMENTO DE ENTIDADES EM MENSAGENS DO TWITTER

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Alberto H. F. Laender
Coorientador: Adriano Veloso

Belo Horizonte
Outubro de 2012

de Oliveira, Diego Marinho

O48a Uma Abordagem Baseada em Fluxo de Filtros para o Reconhecimento de Entidades em Mensagens do Twitter / Diego Marinho de Oliveira. — Belo Horizonte, 2012
xvi, 71 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.
Orientador: Alberto H. F. Laender
Coorientador: Adriano Veloso

1. Computação - Teses. 2. Redes sociais on-line - Teses. 3. Twitter. - Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma abordagem baseada em fluxo de filtros para o reconhecimento de entidades em mensagens do twitter

DIEGO MARINHO DE OLIVEIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO ALONSO VELOSO - Co-orientador
Departamento de Ciência da Computação - UFMG

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

PROF. LUIS ENRIQUE ZÁRATE
Departamento de Ciência da Computação - PUC-MG

PROF. RENATO MARTINS ASSUNÇÃO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de outubro de 2012.

Agradecimentos

Agradeço primeiramente a Deus o dom da vida, a graça dessa conquista e a humildade que me permitiu seguir em frente mesmo nos momentos mais difíceis. Sem Ele não chegaria onde estou.

Aos meus pais, exemplos de conduta, que me deram a oportunidade de estudar, escolher os meus caminhos e a compreensão que me reservaram nesse momento de abdicações em prol do meu aprimoramento intelectual, a minha gratidão. Os sentimentos nunca em mim faltaram, apenas foram silenciosos, mas explícitos na certeza de meus esforços para que eu pudesse ser um pouco do tanto que merecem. À minha família, especialmente às minhas irmãs e aos meus avós, cada um a sua maneira, que me incentivou nessa caminhada.

À Maiza que esteve ao meu lado em todos os momentos, agradeço a paciência, carinho e confiança. Houve momentos em que fui bastante ausente e, por isso, peço desculpas e agradeço por ter essa companheira maravilhosa.

Ao meu professor orientador doutor Alberto Laender, que com a sua inteligência, simplicidade, confiança, e, principalmente, paciência, que me possibilitou concluir essa etapa tão importante de minha vida e torná-la menos árdua. Ao meu coorientador Adriano Veloso que por meio dos seus conhecimentos matemáticos foi de fundamental importância na colaboração da parte técnica da abordagem proposta nesta dissertação.

Aos meus colegas do LBD, especialmente, ao Hasan, Thiago Sales, Thiago Cardoso, Péterson e Allan Jones pelo companheirismo e momentos de descontração. Aos meus professores de graduação, especialmente, aos meus orientadores de pesquisa científica doutor Zárate e doutora Cristiane, que por meio dos seus ensinamentos me deram a base para a consolidação de um sonho. À todos que de alguma forma contribuíram ou torceram por mim, meu reconhecimento!

"A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo."
(Albert Einstein)

Resumo

A tarefa de reconhecimento de entidades consiste em se localizar e classificar elementos em um texto não estruturado por meio de técnicas de processamento de linguagem natural apropriadas ao domínio da aplicação. No contexto da Web, essa tarefa é fundamental para a identificação de entidades, tais como pessoas, organizações, lugares, entre outras. Recentemente, microblogs como o *Twitter* e o *Tumblr* tornaram-se um fenômeno na Web, representando um novo desafio para o reconhecimento de entidades. No *Twitter*, por exemplo, trafega um grande volume de mensagens em um curto espaço de tempo, dificultando essa tarefa e a extração de informação sobre um determinado assunto. Além disso, o ambiente do *Twitter* é bastante dinâmico e orientado a fluxo de dados, necessitando, assim, de ferramentas e métodos adequados às suas características. Não há na literatura, no entanto, muitos trabalhos que tratam desse assunto, evidenciando uma ampla área de pesquisa a ser realizada para o reconhecimento de entidades nesse ambiente. Dessa forma, esta dissertação propõe uma abordagem alternativa denominada FS-NER (do inglês *Filter Stream Named Entity Recognition*) para a realização dessa tarefa. A abordagem FS-NER se baseia na utilização de filtros de forma independente e rápida, altamente escalável e adequada ao ambiente do *Twitter* para o reconhecimento de entidades. A fim de avaliar a eficácia da abordagem proposta, realizou-se um conjunto exaustivo de experimentos utilizando-se mensagens do *Twitter*. Nesses experimentos, foram empregadas três coleções distintas: uma contendo mensagens em inglês, outra em português e a terceira em idiomas diversos. Os resultados obtidos demonstraram que apesar da simplicidades dos filtros usados, a abordagem proposta foi capaz de superar as outras baseadas em *Conditional Random Fields* com melhoria média de 3% para a métrica F_1. Além disso, essa abordagem apresenta ordem de magnitude mais rápida e, portanto, mais apropriada para o paradigma de fluxo de dados típico do *Twitter*.

Palavras-chave: Reconhecimento de Entidades, *Conditional Random Fields*, Redes Sociais, *Microblogs*, *Twitter*.

Abstract

The task of entity named recognition is to locate and classify elements in unstructured text through techniques of natural language processing appropriate to the application domain. In the Web context, this task is critical to the identification of entities such as people, organizations, places, among others. Recently, microblogs like Twitter and Tumblr became a phenomenon on the Web, representing a new challenge for the recognition of entities. In Twitter, for example, traffic a large volume of messages in a short time, difficulting the task and the extraction of information about a particular subject. Moreover, the Twitter environment is quite dynamic and driven by data stream, requiring thus tools and methods suited to its characteristics. There is not in the literature, however, many works that address this issue, showing a wide area of ãâresearch to be conducted for named entity recognition in this environment. Thus, this master thesis proposes an alternative approach to perform this task called FS-NER (Filter Stream Named Entity Recognition). The FS-NER approach is based on the use of filters in an independent and fast manner, highly scalable and suitable for the environment of the Twitter for named entity recognition. In order to evaluate the effectiveness of the proposed approach, we carried out an exhaustive set of experiments using messages of Twitter. In these experiments, we used three distinct collections: one containing messages in English, one in Portuguese and third in several languages. The results showed that despite the simplicities of the filters used, the proposed approach was able to outperform the other approach based on Conditional Random Fields with improvement mean of 3% for the F_1 metric. Moreover, this approach presents orders of magnitude faster and therefore more suitable for the typical data stream paradigm of Twitter.

Keywords: Named Entity Recognition, Conditional Random Fields, Social Networks, Microblogs, Twitter.

Sumário

Capítulo 1

Introdução

1.1 Introdução

O reconhecimento de entidades (do inglês *Named Entity Recognition* - NER) é uma tarefa típica de extração de informação [Ekbal et al., 2010a]. Localizar e classificar elementos em um texto não estruturado requer o uso de técnicas apropriadas ao domínio da aplicação [Kazama et al., 2002; Paşca, 2004, 2007; Ruch et al., 2003; Tanabe et al., 2005]. No contexto da Web, essa tarefa é fundamental para a identificação de entidades, tais como pessoas, organizações, lugares, entre outras de interesse em determinadas aplicações que fazem uso de textos não estruturados [Nadeau & Sekine, 2007].

Recentemente, microblogs como o *Twitter*[1] e o *Tumblr*[2] tornaram-se um fenômeno na Web, representando um novo desafio para o reconhecimento de entidades. No *Twitter*, por exemplo, trafega um grande volume de mensagens em um curto espaço de tempo, dificultando a tarefa de reconhecimento e extração de informação sobre determinado assunto [Ediger et al., 2010; Noordhuis et al., 2010]. Essas mensagens também não possuem qualquer formatação e apresentam grande variação na escrita para a mesma semântica (e.g., "você" possui variações como "vc" ou simplesmente "c"), além de conterem erros ortográficos. Além disso, as mensagens muitas vezes incluem palavras em vários idiomas que podem adotar sinais ou códigos específicos de um grupo de pessoas em um determinado contexto (e.g., gírias, abreviações não vigentes no idioma local, entre outros). Por esses e outros fatores, como o tamanho reduzido das mensagens postadas, torna-se mais desafiador o reconhecimento de entidades em ambientes como o *Twitter*.

Nos últimos dois anos, devido ao sucesso das redes sociais, alguns trabalhos sur-

[1]Disponível em `http://www.twitter.com`
[2]Disponível em `http://www.tumblr.com`

giram com a intenção de investigar como realizar o reconhecimento de entidades nesse meio pouco explorado, no qual técnicas apropriadas ainda não foram consolidadas [Michelson & Macskassy, 2010; Liu et al., 2011; Ritter et al., 2011]. Nos trabalhos de Ritter et al. [2011] e Gimpel et al. [2011], por exemplo, os autores destacam que as técnicas tradicionais adotadas até então em várias das tarefas de reconhecimento de entidades não podem ser empregadas diretamente no ambiente das redes sociais.

Dessa forma, esta dissertação propõe uma abordagem alternativa para o reconhecimento de entidades denominada FS-NER (do inglês *Filter Stream Named Entity Recognition*) que é mais adequada para lidar com mensagens do *Twitter* do que as técnicas tradicionais. Essencialmente, o processo de reconhecimento de entidades pode ser visto como um fluxo de grande volume de mensagens do *Twitter* (i.e., fluxo de *tweets*) controladas por uma série de componentes, denominados *filtros*. Um filtro recebe uma mensagem transmitida na forma de fluxo de dados, realiza processamentos específicos nesta mensagem e retorna a informação sobre as possíveis entidades nela contidas (i.e., cada filtro é responsável por reconhecer entidades de acordo com algum critério específico). Filtros podem ser simples a ponto de considerar somente letras maiúsculas ou usar dicionários de termos para o reconhecimento de entidades. Mas caso necessário, os filtros podem ser combinados, linearmente ou não, melhorando drasticamente o resultado final da tarefa de reconhecimento de entidades por explorar a independência e complementaridade dos diversos filtros.

Para avaliar a eficácia da abordagem proposta nesta dissertação, realizou-se um conjunto de experimentos usando dados do *Twitter*. Empregou-se nesses experimentos três coleções distintas: uma contendo mensagens em inglês, outra contendo mensagens em português e uma terceira contendo mensagens em idiomas diversos. As avaliações se basearam na identificação de sete tipos de entidade, em que os resultados obtidos pela abordagem FS-NER foram comparados com os resultados obtidos por abordagens tradicionais baseadas em *Conditional Random Fields* [Lafferty et al., 2001], que são consideradas o estado da arte para a realização desta tarefa. A partir dos resultados obtidos, como será relatado em mais detalhes nesta dissertação, foi possível mostrar que, apesar da simplicidades dos filtros usados, a abordagem proposta foi capaz de superar as abordagens tradicionais com melhoria média de 3%, além de ser ordens de magnitude mais rápida e, assim, mais apropriada para o paradigma de fluxo de dados típico do *Twitter*.

1.2 Contextualização e Motivação

Inicialmente introduzida em 1995 na sexta edição da MUC (acrônimo do inglês *Message Understanding Conference*), a tarefa de reconhecimento de entidades visa o reconhecimento e a extração de informação sobre determinado grupo de termos denominados entidades. Essa tarefa tem por objetivo encontrar termos semelhantes em documentos, com a menor intervenção humana possível. Dessa forma, entende-se por entidades quaisquer termos que se tenha interesse em reconhecer quando se analisa um conjunto de documentos. Exemplos de tipos de entidade que podem ser citados são nomes de pessoas, organizações e locais, determinados valores numéricos (e.g., datas, valor monetário, telefone), entre outros. Tipicamente, o reconhecimento de entidades é aplicado quando se deseja identificar entidades em documentos contendo notícias [Tjong & Erik, 2002; Ekbal & Bandyopadhyay, 2008; Stern & Sagot, 2010], páginas semi-estruturadas encontradas em sítios da Web [Downey et al., 2007; Whitelaw et al., 2008; Zhu, 2009], textos da área biomédica [Ekbal et al., 2010b; Sætre et al., 2010], entre outros casos. Entretanto, não é somente da aplicação direta do reconhecimento de entidades que se obtém os benefícios desse processo. Assim, o reconhecimento de entidades pode ser usado como um processo auxiliar em tarefas mais complexas. Pode-se aplicar o reconhecimento de entidades para realizar a mineração de dados na Web [Jiang, 2012], monitorar eventos [Mesquita et al., 2010; White et al., 2010; Asur & Huberman, 2010; Oh et al., 2011], analisar opiniões dos usuários [Gînscă et al., 2011], entre outras aplicações que precisem da informação relacionada às entidades.

Apesar de o reconhecimento de entidades ser geralmente empregado em ambientes mais formais como exemplificado acima, *microblogs* como o *Twitter* e o *Tumblr* estão transformando a forma como as pessoas se comunicam, possibilitando que novas fontes de informação sejam exploradas. Recentemente, com o surgimento e expansão das redes sociais, tornou-se de interesse analisar esse tipo de contexto em busca de informação relevante ao usuário. Estima-se que plataformas importantes como o *Twitter* possuem mais de 500 milhões de usuários que geram mais de 340 milhões de mensagens diariamente, produzindo, desta forma, uma fonte de dados única para pesquisa sobre a Web.

Entretanto, devido às características específicas das mensagens publicadas nas redes sociais, o reconhecimento de entidades nesse tipo de ambiente é considerado bastante informal, diferenciando drasticamente dos ambientes mais tradicionais. Trabalhos recentes [Liu et al., 2011; Locke & Martin, 2009; Ritter et al., 2011] relatam várias dificuldades e impedimentos em aplicar técnicas tradicionais de reconhecimento de entidades em dados do *Twitter*, sugerindo desta forma a necessidade de ferramen-

tas e métodos mais flexíveis e efetivos para cuidar dessa tarefa em contextos mais desafiadores.

A seguir, são listados os principais desafios existentes quando se considera a tarefa de reconhecimento de entidades em dados do *Twitter*. A saber, os desafios discutidos são relacionados ao grande volume de dados, falta de formalismo, dependência do idioma, ambiente dinâmico, falta de contextualização das mensagens e orientação a fluxos de dados.

Grande volume de dados. O *Twitter* produz um grande volume de dados todos os dias devido ao grande número de usuários e a intensa interação entre eles. Isso significa que métodos e ferramentas mais eficientes são necessários para lidar com o reconhecimento de entidades em dados do *Twitter*. Por exemplo, abordagens que requerem um processo iterativo para gerar seus modelos de reconhecimento de entidades podem ter seu desempenho seriamente afetados devido a demora na convergência de seus parâmetros [Sha & Pereira, 2003]. Considerando cenários reais, o uso de tais abordagens pode se tornar um gargalo em termos de desempenho computacional. Abordagens probabilísticas que confiam em processos de aprendizado iterativo devem utilizar-se de características mais leves e eficientes para possibilitar o reconhecimento de entidades nesse ambiente.

Falta de formalismo. As redes sociais, assim como o *Twitter*, em geral são ambientes em que se predomina a informalidade textual. A informalidade é capaz de prejudicar o desempenho das ferramentas no reconhecimento de entidades devido à existência de casos frequentes em que as mensagens estão fora da norma culta do idioma, além de apresentarem bastantes erros ortográficos. Considerando as técnicas tradicionais de reconhecimento de entidades, as mesmas se tornam inapropriadas em ambientes como esse e, portanto, adaptações se tornam necessárias para amenizar os problemas de informalidade. A Figura 1.1 apresenta o exemplo de uma mensagem em que muito desses problemas estão presentes. Os números na figura apontam detalhes que dificultam o processo de reconhecimento de entidades.

Em (1), o nome próprio inicia com letra em minúscula. Em geral, termos iniciados com letra maiúscula constituem forte evidência textual para auxiliar na identificação de entidades. Recorrentemente, no *Twitter* esse é um dos principais problemas a se lidar. Há outros casos, em que letras em maiúscula são usadas arbitrariamente em posições diferentes ao início do termo, além de serem adotadas indistintamente para termos candidatos a entidade ou não (e.g., “CHELSEA”, “ChelSEA”, “TALKS”, etc); (2) o verbo não é flexionado de acordo com a terceira pessoa; (3) há presença de er-

ros ortográficos. Em várias situações os erros ortográficos são derivados das técnicas de *texting*, por exemplo "*2nd Nov: 4 Tet b2b Caribou DJ set + loads m0r3.*"; (4) pontuação errada e falta de padronização podem causar ambiguidade, dificultado o reconhecimento de entidades a partir das evidências textuais usadas, além de provocar a segmentação da mensagem inapropriadamente; (5) apresenta *hiperlinks* no meio do texto fragmentando o sentido da mensagem. Em geral, os *hiperlinks* são simplificados gerando *hiperlinks* quebrados, que talvez levem a páginas que não estão relacionadas ao conteúdo da mensagem; (6) novamente o aparecimento de pontuação equivocada aumenta a ambiguidade do texto; (7) o emprego incorreto de pontuação próxima a uma palavra, dificulta o uso de técnicas convencionais de reconhecimento de entidades, pois a palavra seguida de ponto deveria ser iniciada por letra maiúscula; (8) presença fora de contexto de símbolos especiais do *Twitter* denominados *hashtags*. Esses termos, quando mal empregados, podem deteriorar o processo de reconhecimento, descaracterizando a estrutura da mensagem; (9) a presença de erros anteriores (4, 5 e 6) provocam a segmentação da mensagem em duas partes, as quais perdem sentido se não forem analisadas conjuntamente, podendo provocar a degradação no processo de reconhecimento. O exemplo da Figura 1.1 apresenta algumas situações que ilustram como a falta de formalismo pode representar um desafio para as abordagens típicas de reconhecimento de entidades quando aplicadas a mensagens do *Twitter*.

Figura 1.1: Exemplo de uma mensagem informal apresentando nove aspectos típicos de informalismo.

Diversidade de idiomas. Apesar da predominância de alguns idiomas, tais como Inglês, Japonês, Português e Espanhol, o Twitter apresenta uma enorme diversidade nesse aspecto [Hong et al., 2011]. Um desafio particular ocorre quando é necessário identificar entidades em idiomas, tais como o Bengali [Ekbal et al., 2010a], no qual o reconhecimento de entidades é intrinsecamente mais complexo devido as especificidades gramaticais do idioma. Além disso, a necessidade de processar diferentes idiomas pode introduzir dificuldades na realização dessa tarefa no Twitter, sendo que abordagens que dependem excessivamente de características definidas para um idioma podem se tornar

inadequadas nesse ambiente. Outro problema desafiador que deve ser considerado é quando idiomas distintos se sobrepõem em um mesmo domínio observado. Mesmo quando não se considera diferentes idiomas, é possível haver grandes discrepâncias de escritas devido a valores culturais. Desta forma, a necessidade de processar diferentes idiomas pode introduzir dificuldades no processo de reconhecimento de entidades e abordagens que dependem excessivamente de evidências textuais específicas de um idioma podem tornar-se inadequadas.

Ambiente dinâmico. O *Twitter* caracteriza-se por ser um ambiente bastante dinâmico. Em poucos segundos uma grande quantidade de mensagens é trocada entre os seus usuários. Essa constante e rápida interação provoca a mudança dos assuntos discutidos na rede, aumentando ainda mais os desafios que devem ser enfrentados. Dessa forma, o principal desafio está relacionado à falta de robustez das ferramentas e métodos para acompanhar as mudanças dos assuntos discutidos nas mensagens do *Twitter*. Muitas das técnicas empregadas para o reconhecimento de entidades são custosas tanto em desempenho quanto em complexidade e, portanto, não são adequadas a esse ambiente. Quando são adaptáveis, em geral há um alto custo computacional para acompanhar as mudanças de contexto dos assuntos discutidos pelos usuários. Também, as técnicas adotadas devem ser capazes de lidar com a falta de confiabilidade e ambiguidade das mensagens trafegadas.

Em geral, a consequência da falta de habilidade em lidar com ambientes dinâmicos como o *Twitter* é a degradação das ferramentas até se tornarem inaptas para realizar o reconhecimento. Em muitas das vezes, a degradação ocorre principalmente por três razões: (i) aparição de novas entidades; (ii) aparição de novos contextos e situações em que as entidades são mencionadas; (iii) utilização de mensagens inadequadamente rotuladas. Solucionar o problema para qualquer um desses casos é uma tarefa não trivial que contribui para o progresso da área.

Falta de contextualização dos dados. As imposições de tamanho das mensagens publicadas no *Twitter*, limitadas a 140 caracteres, podem reduzir a eficiência na realização do reconhecimento de entidades. Os desafios associados a essa limitação estão relacionados à falta de evidências textuais suficientes para o reconhecimento considerando o contexto reduzido. Como exemplo, suponha que se deseja reconhecer entidades do tipo *Companhia* na mensagem *"RT: I bought at J&J."*. Considerando os termos dessa mensagem, *"J&J"* é avaliado como termo candidato. Entretanto, somente com a informação disponível, reconhecer *"J&J"* como nome de uma companhia pode levar a equívocos. Se por acaso houvesse outra informação adicional, tal como *"In which **x** did*

you find it?" a ambiguidade poderia ser resolvida. Se $x =$ "*city*" então possivelmente poderia-se descartar *"J&J"* como entidade. Em outra situação, caso $x =$ "*store*" então *"J&J"* poderia ser avaliada como entidade. Apesar de ser um exemplo ingênuo, ele demostra que a contextualização da mensagem auxilia no processo de reconhecimento. Não obstante, situações como essa ocorrem com frequência no *Twitter* e, portanto, ao realizar o reconhecimento de entidades nesse ambiente, aspectos como esse devem ser levados em consideração.

Orientação a fluxo de dados. O *Twitter* é caracterizado pela transmissão de mensagens na forma de fluxo de dados. Esse tipo de transmissão faz com que as mensagens sejam rapidamente disseminados pela rede. Logo, torna-se essencial projetar soluções de reconhecimento de entidades que levem em consideração o rápido surgimento de novas informações. Além disso, somado aos desafios discutidos sobre a natureza dinâmica do *Twitter*, tem-se um outro grande desafio a ser enfrentado para realizar o reconhecimento de entidades adequadamente. Em geral, ferramentas e métodos voltados para o reconhecimento de entidades tratam os dados na forma de conjuntos (i.e., arquivos em lote), ao invés de tratá-los na forma de fluxos como é necessário no ambiente do *Twitter*. A mudança nesse paradigma implica em novos desafios que devem ser enfrentados para o reconhecimento nesse tipo de ambiente. Logo, os principais desafios estão relacionados à criação ou adaptação de abordagens capazes de se utilizarem do fluxo de dados para gerar e atualizar os modelos estatísticos e probabilísticos empregados para o reconhecimento de entidades a um baixo custo computacional.

Por fim, os desafios supracitados sintetizam algumas das principais dificuldades encontradas na realização do reconhecimento de entidades no ambiente do *Twitter*. Logo, vê-se como motivação para esta dissertação atacar e amenizar os problemas de reconhecimento de entidades relacionados à esse tipo de ambiente. Além disso, sabe-se que no contexto das redes sociais, o reconhecimento de entidades ainda é pouco explorado e novas alternativas tornam-se necessárias para consolidação de métodos e ferramentas eficazes para essa tarefa.

Dentre os vários tipos de aplicação e serviços que podem se beneficiar do reconhecimento de entidades em redes sociais, podem ser citados os conhecidos como MMS (Monitoramento de Mídia Social). Os serviços de MMS podem se utilizar das informações publicadas pelos usuários com o objetivo de informar mais efetivamente as tendências de um determinado evento observado [Asur & Huberman, 2010]. Serviços

como o *Observatório da Web*[3] e o *Google Dengue Trends*[4] consistem em informar em tempo real o cenário atual sobre um determinado assunto. Serviços como esses, em geral, podem se beneficiar do uso do reconhecimento de entidades. Portanto, torna-se bastante relevante o desenvolvimento de abordagens mais eficientes para a rápida e eficiente entrega de informação para consumo desses serviços.

1.3 Objetivo

O objetivo desta dissertação é apresentar uma nova abordagem para o reconhecimento de entidades em mensagens publicadas no *Twitter*. Especificamente, pretendemos descrever essa abordagem e consolidá-la apresentado os resultados de experimentos realizados que a comparam com abordagens comumente adotadas para solucionar esse tipo de problema. Embasamos essa abordagem a partir da experiência adquirida na utilização de arcabouços baseados em *Conditional Random Fields* [Lafferty et al., 2001], de forma a empregar algoritmos simples e eficientes para auxiliar o processo de reconhecimento.

1.4 Contribuições

As contribuições desta dissertação são:

- **Abordagem inovadora para o reconhecimento de entidades em mensagens do *Twitter*.** Proposta de uma nova abordagem baseada em filtros para tratar vários dos problemas citados envolvendo o reconhecimento de entidades em mensagens do *Twitter*. A partir da observação de que filtros poderiam utilizar-se das evidências textuais empregadas no processo de reconhecimento, tornou-se possível elaborar mecanismos mais simples e adequados ao problema, e que fossem capazes de executar a mesma tarefa satisfatoriamente.

- **Exaustiva avaliação experimental.** Realizou-se uma exaustiva avaliação experimental analisando-se as principais estratégias para a combinação dos filtros adotados na abordagem proposta. Além disso, comparou-se a abordagem proposta com arcabouços baseados em *Conditional Random Fields*, que são considerados o estado da arte para a realização desta tarefa.

[3] Observatório da Web é um projeto de pesquisa pioneiro do InWeb-Instituto Nacional de Ciência e Tecnologia para a Web que está sendo desenvolvido pelo Departamento de Ciência da Computação da Universidade Federal de Minas Gerais (UFMG). Disponível em `http://observatorio.inweb.org.br/`

[4] Disponível em `http://www.google.org/denguetrends/`.

- **Compilação de boas práticas para o reconhecimento de entidades no *Twitter*.** Considerando as várias dificuldades existentes para realizar o reconhecimento de entidades em mensagens do *Twitter*, nesta dissertação é compilado um conjunto importante de boas práticas para a realização dessa tarefa. Essas boas práticas foram adquiridas empiricamente e são repassadas como forma construtiva de facilitar a qualquer pesquisador ter conhecimento sobre observações essenciais e que são raramente relatadas por outros trabalhos realizados na área.

1.5 Organização da Dissertação

O restante desta dissertação está organizado da seguinte forma. O Capítulo 2 apresenta os fundamentos e trabalhos relacionados ao tema abordado. Os fundamentos apresentados têm como objetivo rever os conceitos necessários e exigidos para que se tenha um entendimento completo do trabalho desenvolvido. Os trabalhos relacionados estão divididos em duas seções, abordando os trabalhos de reconhecimento de entidades em geral e aqueles especificamente voltados para as redes sociais. O Capítulo 3 descreve e formaliza a abordagem proposta denominada FS-NER (acrônimo em inglês para *Filter Stream Named Entity Recognition*). Além da apresentação da abordagem são introduzidos os cinco filtros básicos usados na avaliação experimental realizada. O Capítulo 4 descreve detalhadamente a avaliação experimental da abordagem proposta, apresentando os principais resultados obtidos. Por último, o Capítulo 5 apresenta as conclusões e os trabalhos futuros.

Capítulo 2

Fundamentos e Trabalhos Relacionados

2.1 Conceitos Básicos e Notações

Esta seção apresenta os principais conceitos básicos empregados nesta dissertação. Tais conceitos, em sua maioria, baseiam-se nos principais trabalhos da literatura relacionados ao reconhecimento de entidades.

Termo. Denomina-se termo todo conjunto de caracteres que não possua espaço em branco. Assim, considera-se como termo qualquer sequência de caracteres, incluindo pontuações representada por x. Como exemplos de termos podemos citar *"Pesquisa"*, *"25/12"*, *"10h15"*, *"http://dcc.ufmg.br/pos"*, *"gooooool!!"*, *"@user"* e *"#@!25..."*.

Sequência de Termos. Denomina-se sequência de termos (ou sentença) todo conjunto de termos em que cada termo possui uma posição i correspondente na sequência. Por convenção, os termos em uma sequência são separados por espaço em branco. Uma sequência X formada por vários termos x_i é representada, então, por $\{x_i | x_i \in X\}$. Desse modo, uma sequência X pode ser dada por $X = \{$*"O"*, *"computador"*, *"Watson"*, *"da"*, *"IBM"*, *"venceu"*, *"os"*, *"dois"*, *"maiores"*, *"campeões"*, *"do"*, *"Jeopardy"*$\}$, em que os termos nas posições 3 e 5 da sequência são $x_3 =$ *"Watson"* e $x_5 =$ *"IBM"*.

Rótulo. Denomina-se rótulo todo valor associado ao termo x na posição i de uma sequência. Em geral, os rótulos são importantes para delimitar os termos correspondentes às entidades. Além disso, os rótulos podem indicar algumas informações de

interesses, entre elas, classes gramaticais e estado do termo. Desse modo, um rótulo y recebe um dos possíveis valores existentes em um alfabeto preestabelecido $\mathcal{Y}$. O alfabeto de rótulos[1] $\mathcal{Y} = \{Entidade, Outro\}$, por exemplo, apresenta $y_5 = Entidade$ para o rótulo na posição 5.

Sequência de Rótulos. Denomina-se sequência de rótulos todo conjunto de rótulos Y em que cada rótulo y possui uma posição i correspondente na sequência. Cada rótulo y deve estar associado a pelo menos um termo x, de modo que essa associação é definida por (X, Y). Uma sequência Y formada por vários rótulos y, então, é dada por $\{y_i | y_i \in Y\}$. Ao considerar a sequência X retratada anteriormente e supondo-se que se deseja rotular entidades do tipo *Organização* e *Programa de Televisão*, o resultado, é dado por Y =*{Outro, Outro, Outro, Outro, Outro, Entidade, Outro, Outro, Outro, Outro, Outro, Outro, Outro, Entidade}*.

Entidade. Denomina-se entidade qualquer termo de interesse para reconhecimento nos conjuntos de documentos em análise. Consideram-se, assim, os termos cujos descritores estejam de acordo com algum critério preestabelecido como entidade [Kripke, 1980]. Os descritores, por exemplo, podem considerar os critérios como nomes de pessoas associadas ao cargo de chefe de estado para restringir o tipo de entidade *Chefe de Estado* na sequência representada abaixo.

> "*Em Cannes, na França, para participar da Cúpula do G20, a presidenta Dilma Rousseff agendou para quarta uma série de reuniões. No primeiro compromisso do dia, Dilma conversa com a primeira-ministra da Austrália, Julia Gillard. Depois, reúne-se com o diretor-geral da Organização Internacional do Trabalho, Juan Somavia, e depois com o presidente da República Popular da China, Hu Jintao.*"

Desse modo, o conjunto de termos "*Juan Somavia*" não é uma entidade, enquanto "*Dilma Rousseff*", "*Dilma*", "*Julia Gillard*" e "*Hu Jintao*" se tratam de entidades. Os tipos de entidade podem, também, ser representados por valores numéricos, datas, entre outros.

Em geral, os termos considerados como pertencentes aos tipos de entidade são envolvidos por delimitadores. Inicialmente, nos primórdios da tarefa de reconhecimento de entidades, os primeiros delimitadores foram denominados ENAMEX, NUMEX e TIMEX. Os ENAMEX, por exemplo, desenvolvidos a partir de uma das edições da MUC, foram originalmente criados para delimitar entidades do tipo pessoa (PER) e

[1] Os rótulos serão detalhadamente discutidos na Seção 2.3.

organização (ORG) e posteriormente estendidos para outros tipos [Grishman & Sundheim, 1996]. A partir da definição inicial e devido à demanda de reconhecimento de diversos tipos de entidade, hoje há centenas de delimitadores distintos que podem ser adotados para destacar as entidades de interesse em um documento. Sekine & Nobata [2004], por exemplo, propõem cerca de 150 delimitadores para entidades.

2.2 Reconhecimento de Entidades

Conforme já mencionado, o reconhecimento de entidades foi inicialmente introduzido como parte da MUC-6 em 1995 [Grishman & Sundheim, 1996] e consiste em uma subtarefa da extração de informação [Ekbal et al., 2010a]. O objetivo principal é o reconhecimento de entidades em quaisquer tipos de documento ou fragmento de texto. Nesse sentido, dado um termo, busca-se identificar qual é o rótulo correspondente, podendo ser esse rótulo um valor que indique se o termo é entidade ou não. Identificar entidades em um texto não estruturado torna-se complexo e, portanto, vários métodos e ferramentas [Kazama et al., 2002; Paşca, 2004, 2007; Riaz, 2010; Ruch et al., 2003; Tanabe et al., 2005] têm sido desenvolvidos.

A Figura 2.1 apresenta um possível resultado para o reconhecimento de entidades que considerou nomes de *Pessoas* (PER), *Organizações* (ORG), *Locais* (LOC) e *Outros* (MISC) como entidades de interesse. Pode-se notar, assim, a delimitação dos termos selecionados como entidades.

> Na avaliação de ***[Per** Dilma Rousseff]*, o ***[Org** governo federal]* considera "essenciais" parcerias com o setor privado para alavancar o crescimento do ***[Loc** Brasil]*. Em relação às medidas tomadas, o governador de ***[Loc** São Paulo]*, ***[Per** Geraldo Alckmin]* (***[Misc** PSDB]*), disse nesta ***[Misc** quarta-feira]* (15) que o modelo de concessões, anunciado pelo ***[Org** governo federal]*, é uma "medida correta" e terá todo o seu apoio.

Figura 2.1: Exemplo de um resultado típico da tarefa de reconhecimento de entidades.

2.3 Codificação

A codificação consiste em definir um alfabeto $\mathcal{Y}$ de rótulos específicos de modo que os rótulos pertencentes a esse alfabeto possam ser associados durante o processo de reconhecimento de entidades para cada termo ou conjunto de termos de uma sequência

analisada. Esses rótulos são importantes para delimitar aspectos do texto que serão utilizados por alguma ferramenta ou método responsável em realizar o reconhecimento de entidades. A escolha de uma codificação é um processo importante, pois, essa define a segmentação dos termos e possibilita a utilização das informação geradas por esse processo pelos métodos e ferramentas para o reconhecimento de entidades. Não há na literatura um consenso sobre qual a melhor codificação a ser adotada. Algumas codificações, no entanto, se tornam mais apropriadas à tarefa de reconhecimento de entidades devido a suas propriedades serem mais adequadas ao tipo de entidade e contexto em que são empregadas.

As codificações mais utilizadas são IO, BIO e BILOU, em que cada letra pertencente ao nome da codificação representa um rótulo. Devido à falta de convenções na área de reconhecimento de entidades, muitas das vezes a mesma codificação recebe nomes distintos. Em alguns trabalhos, por exemplo, IO e BIO são listados invertidos (OI e IOB). O caso mais comum de falta de convenção é para a codificação BILOU. Em muitos trabalhos essa aparece como IOBEW [Leaman & Gonzalez, 2008] e BCEUO [Sarawagi, 2006]. De acordo com as informações apresentadas, define-se abaixo as três codificações e que, por questões de nomenclatura, serão mantidas as denominações no idioma inglês.

IO. A codificação IO é definida por $\mathcal{Y} = \{Inside,\ Outside\}$ e representa a opção mais simples de codificação. Associa-se o rótulo *Inside* ao termo x_i quando o mesmo é entidade ou *Outside* caso contrário. Apesar dessa codificação ser simples e direta, a mesma não é capaz de informar o começo e o fim de entidades adjacentes em um texto, como termos de entidades distintos que estão separados por uma posição. Isso ocorre devido a cada entidade receber o mesmo rótulo, que no caso é *Inside*.

BIO. A codificação BIO é definida por $\mathcal{Y} = \{Beginning,\ Inside,\ Outside\}$. Devido à adição do rótulo *Beginning*, a codificação BIO soluciona o problema enfrentado pela codificação IO. No caso, esse rótulo é adicionado sempre à primeira posição de um termo de entidade distinta. Esse rótulo evidência o começo de uma nova entidade.

BILOU. A codificação BILOU é definida por $\mathcal{Y} = \{Beginning,\ Inside,\ Last,\ Outside,\ UnitToken\}$. Em particular, essa codificação separa as entidades em entidades simples, formadas por um único termo, e entidades compostas, formadas por dois ou mais termos. Desse modo, uma entidade simples é rotulada com *UnitToken*, enquanto entidades compostas recebem os rótulos *Beginning*, *Inside* e *Last*. No caso de uma entidade composta, são adotados os rótulos *Beginning* e *Last* para delimitar, respecti-

vamente, o começo e o final da entidade, e *Inside* para as posições intermediárias da entidade.

A Figura 2.2 ilustra exemplos distintos para a codificação BILOU. No Exemplo 1 é apresentada a rotulação de uma entidade quando a mesma é composta de um único termo. Nesse caso, usa-se o rótulo *UnitToken* (U). No Exemplo 2, a entidade é composta por dois termos e, por isso, o primeiro recebe o rótulo *Beginning* (B) e o segundo recebe o rótulo *Last* (L). Finalmente, no Exemplo 3, como a entidade é composta por mais de dois termos, o primeiro e o último termos recebem respectivamente os rótulos *Beginning* e *Last*, enquanto que o termo intermediário recebe o rótulo *Inside* (I).

Figura 2.2: Exemplos para a codificação BILOU.

Exemplo 1: Entidade simples								
U	O	O	O	O	O	O	O	O
Obama	sancionou	o	aumento	do	teto	das	dívidas	(terça)

Exemplo 2: Entidade composta com tamanho dois.								
O	O	O	O	O	O	O	*B*	**L**
Mais	um	tremor	ocorreu	pela	manhã	no	**arquipélago**	**japonês**

Exemplo 3: Entidade composta com tamanho três.								
O	**B**	**I**	**L**	O	B	L	O	U
Grupo	**Pão**	**de**	**Açúcar**	compra	Casas	Bahia	pela	*Globex*

As três codificações apresentam diferentes características como simplicidade e boa capacidade de fornecer informação. Devido às avaliações experimentais relatadas por Ratinov & Roth [2009], optou-se nesta dissertação pelo o uso da codificação BILOU nos experimentos realizados.

2.4 Conjunto de Características para o Reconhecimento de Entidades

O conjunto de características (*features*) para o reconhecimento de entidades representa um mecanismo auxiliar na determinação de um rótulo y_i a partir de um termo x_i, em que cada característica corresponde a uma função binária que produz o valor *1* quando a função é válida ou *0* caso contrário. Uma caraterística deve ser definida através de uma regra clara que represente uma possível evidência textual relevante sobre os tipos de entidades de interesse. Ao reconhecer, por exemplo, nomes de pessoas, uma caraterística relevante seria retornar 1 se a palavra começar com letra maiúscula ou 0 caso contrário. Essa função torna-se um importante artifício para selecionar as

evidências textuais capazes de produzir informação de interesse para o reconhecimento de entidades.

Devido às particularidades de cada problema, a escolha das evidências é uma atividade fundamental que exige atenção e não deve ser realizada sem qualquer tipo de análise ou espelhada em escolhas que apesar de serem bastante utilizadas, não são adequadas ao tipo de problema. A solução de cada problema torna-se, dessa maneira, dependente das evidências existentes.

Há um número inesgotável de evidências citadas pelos autores atuais, entre eles Nadeau & Sekine [2007] que as dividem em três tipos: termos, listas de termos e documento. Nessa linha de estudo, as evidências de termos referem-se aos detalhes dos termos como afixos, morfologia, pontuações, numerais, classes gramaticais, entre outras. Nesse nível se o termo é um nome próprio, por exemplo, terminado com o sufixo "*Corp.*", então o termo pode ser uma entidade do tipo *Organização*. Quando se considera listas de termos, as características baseiam-se no uso de listas de nomes de entidade, expressões, siglas, dentre outras para inferir novas entidades, sem que para isso as mesmas tenham sido analisadas anteriormente pelas abordagens de reconhecimento de entidade. Por isso, são consideradas bastante úteis, sendo capazes de armazenar grande quantidade de termos ou expressões de interesse para o reconhecimento. Para listas de expressões, por exemplo, considerando-se a expressão *"O Sr.* $\boldsymbol{x}$*"*, o termo substituído em "x" representaria possivelmente uma entidade do tipo *Pessoa*. Por último, quando se considera o documento como um todo, são utilizadas evidências que analisam vários trechos ou conjuntos de documentos, como, por exemplo, características de meta-dados, frequência de palavras sobre vários documentos, análise dos termos em outros contextos, dentre outras que utilizem informação entre documentos. Esses tipos de evidência podem auxiliar consideravelmente o processo de reconhecimento de entidades quando bem empregados. A Seção 3.3 apresenta detalhadamente o conjunto de características utilizadas nesta dissertação.

2.5 Métricas de Avaliação

As métricas de avaliação permitem a comparação dos resultados obtidos pelas abordagens de reconhecimento de entidades e análise do grau de concordância desses resultados com o gabarito gerado por especialistas na tarefa específica. Esse procedimento permite, assim, mensurar o quão próximo está a solução da abordagem para o resultado considerado correto pelos especialistas.

A Figura 2.3 apresenta, no quadro à esquerda, os resultados produzidos por uma

abordagem para o reconhecimento de entidades e, no quadro à direita, o resultado indicado por especialistas. Nesse exemplo, observa-se que algumas entidades foram reconhecidas com o tipo incorreto (e.g., *"Bloomberg"* e *"Zuckerberg"*), outras foram reconhecidas parcialmente (e.g., *"quinta-feira"*, *"US$ 600 milhões"*, *"Mark Zuckerberg"*) e uma não foi reconhecida (e.g., 6%).

*Segundo a agência de notícias [**Misc** Bloomberg], só na [**Misc** quinta]-feira as ações caíram mais de 6%, diminuindo para [**Org** US]$ [**Misc** 600] milhões a fortuna do fundador do [**Org** Facebook], [**Per** Mark] [**Org** Zuckerberg].*	*Segundo a agência de notícias [**Org** Bloomberg], só na [**Misc** quinta-feira] as ações caíram mais de [**Misc** 6%], diminuindo para [**Misc** US$ 600 milhões] a fortuna do fundador do [**Org** Facebook], [**Per** Mark Zuckerberg].*

Figura 2.3: Exemplo de resultados obtidos por uma abordagem para o reconhecimento de entidades (esquerda) e o gabarito produzido por um especialista (direita).

As métricas comumente adotadas para avaliar a qualidade dos resultados obtidos pelas abordagens para o reconhecimento de entidades são precisão (P), revocação (R) e a média harmônica (F_1) [Baeza-Yates & Ribeiro-Neto, 2011]. De forma abrangente, a precisão representa o quão bem a abordagem é capaz de acertar o rótulo dos termos em questão. A revocação representa o quanto dos termos de interesse foi reconhecido. A média harmônia representa o valor combinado entre a precisão e revocação. Em geral, as conferências de reconhecimento de entidades como MUC, ACE (acrônimo em inglês para *Automatic Content Extraction*), CoNLL (acrônimo em Inglês para *Computational Natural Language Learning*), entre outras, possuem formas distintas de se utilizar as métricas e, portanto, cada uma adota sistemas de pontuação distintos. A MUC, por exemplo, possui um sistema de pontuação direto [Chinchor, 1998]. Considera-se três métricas primárias para calcular os valores de precisão, revocação e F_1. A primeira mede o número de respostas corretas (NRC) informadas pela abordagem para o reconhecimento de entidades, a segunda mede o número de identificações realizadas pela abordagem para o reconhecimento de entidades (NAA) e a terceira mede o número total de entidades na solução (NES) apresentada pelos especialistas. Dessa forma, as métricas são calculadas por:

$$P = \frac{NRC}{NAA} \qquad R = \frac{NRC}{NES} \qquad F_1 = 2 \times \frac{P \times R}{P + R}$$

Considerando-se esse sistema de pontuação é possível calcular as métricas cor-

respondentes no exemplo da Figura 2.3. Para cada métrica primária são analisados os respectivos critérios em termos do tipo de entidade, isto é, se a entidade (e.g., *Pessoa*, *Organização*, etc.) e o valor do termo da entidade em observação são os esperados (e.g., esperava-se "*quinta-feira*" e no lugar rotulou-se "*quinta*"). Para o exemplo acima tem-se, então, $NRC = 3$ (valor = 2, tipo = 1), $NAA = 14$ (valor = 7, tipo = 7) e $NES = 12$ (valor = 6, tipo = 6). Logo, os valores para P, R e F_1 correspondem a 0.21, 0.25 e 0.23%, respectivamente.

O CoNLL adota um sistema de pontuação também bastante direto e simples, entretanto, não é capaz de captar os reconhecimentos parciais de entidades nos resultados das métricas. Esse sistema considera como acerto se o termo correspondente for igual ao da resposta e se o termo reconhecido for do mesmo tipo de entidade. No exemplo anterior, o termo *"Bloomberg"* seria, assim, considerado como um erro, ao invés de acerto parcial como no sistema de pontuação MUC apresentado.

O sistema de pontuação ACE, por sua vez, envolve um sistema de avaliação complexo, em que se atribui pesos para realçar determinados tipos de acerto em detrimento de outros. No entanto, apesar desse sistema possibilitar personalização, a análise dos resultados é dificultada.

2.6 Visão Geral das Técnicas de Reconhecimento de Entidades

Os métodos e ferramentas desenvolvidos para realizar o reconhecimento de entidades são baseados em três tipos de abordagem. A primeira é denominada abordagem baseada em regras, a segunda de abordagem estatística e a terceira de abordagem híbrida, envolvendo a união das duas primeiras.

A abordagem baseada em regras se utiliza de critérios predefinidos para a realização do reconhecimento de entidades. Os métodos e ferramentas desenvolvidos, assim, necessitam de conhecimento prévio das particularidades existentes nos tipos de entidade a serem analisados. Em geral, essa abordagem torna-se necessária quando o domínio contém um número finito de características que possam ser mapeadas para um conjunto de regras fundamentais capazes de realizar a tarefa de reconhecimento satisfatoriamente. Desse modo, essa abordagem está indicada em situações em que as entidades apresentem características mais previsíveis, demandem poucas regras específicas e apareçam em trechos de documentos que possuem padrões de escrita bem definidos.

Os métodos e ferramentas da abordagem baseada em regras têm como uma

de suas vantagens serem de imediata aplicação considerando que as regras já foram definidas. No entanto, possuem a desvantagem de serem específicos para um determinado fim, apresentando perda considerável na qualidade dos resultados quando introduzidos novos exemplos no mesmo domínio. Além disso, essa utilização se torna inviável quando deseja-se reconhecer entidades em um domínio complexo, pois as regras podem não ser suficientes ou precisas para realizar o reconhecimento.

Um exemplo de trabalho que envolve o uso de regras para solucionar problemas relacionados ao reconhecimento de entidades é o proposto por Riaz [2010]. Nesse trabalho, o autor propõe o uso de uma abordagem baseada em regras para identificar entidades no idioma Urdu. A principal razão para a adoção dessa abordagem, segundo o autor, é a escassez de informação sobre o idioma, o que impossibilita o uso de técnicas mais sofisticadas que em geral tendem a melhorar os resultados do reconhecimento de entidades (e.g., técnicas de abordagem estatística). Em outro trabalho, Callan & Mitamura [2002] propõem a criação de uma ferramenta a partir da abordagem KENE baseada em regras. KENE tem como premissa a utilização de regras suficientemente genéricas para identificar entidades em casos distintos, ao mesmo tempo que procura obter uma maior precisão na identificação dessas entidades, propriedade, geralmente, ausente nas outras ferramentas.

É possível notar por meio desses trabalhos supracitados que a abordagem baseada em regras é indicada em situações em que não há conjuntos de dados suficientes para realizar cálculos estatísticos e probabilísticos capazes de auxiliar o processo de reconhecimento de entidades de forma adequada. Além disso, essa abordagem não é indicada quando se deseja explorar um domínio mais complexo, em que há um volume muito grande de regras necessárias para serem elaboradas.

A abordagem estatística em geral, diferentemente da abordagem baseada em regras, utiliza-se de técnicas de aprendizado de máquina para realizar a tarefa de reconhecimento de entidades. Isso significa que os métodos e ferramentas dessa abordagem adotam algum processo de aprendizagem a partir de exemplos de padrões e regras existentes de modo a aplicá-los em várias situações distintas sem que para isso seja necessário anteriormente conhecer mais exemplos. Por isso, métodos e ferramentas que seguem a linha de aprendizado de máquina são caracterizados por necessitar da realização do aprendizado utilizando-se conjuntos de dados denominados de treinamento e validação. Nesse caso, os conjuntos de treinamento são usados para que o método aprenda os padrões existentes a partir dos exemplos informados. Já os conjuntos de validação são usados após a etapa de aprendizagem para verificar se os métodos e ferramentas conseguem atribuir a saída correta sem que tenham conhecimento da resposta previamente. Como exemplo, o trabalho realizado por Zhang et al. [2004] utiliza-se de

técnicas de aprendizado de máquina para efetuar o reconhecimento de entidades em notícias. O objetivo principal dos autores é criar resumos gerados automaticamente das notícias analisadas a partir das entidades chave reconhecidas nos textos. Para melhor obtenção dos resultados, os autores adotaram três técnicas de aprendizado de máquina: árvore de decisão [Tsang et al., 2011], *naive bayes* [Jiang et al., 2009] e o método robusto de minimização de risco [Zhang & Johnson, 2003]. Ao realizar os experimentos, o método robusto de minimização de risco obteve o melhor desempenho na execução das tarefas.

Em outro trabalho, Irmak & Kraft [2010] fazem uso de uma técnica de aprendizado de máquina para melhorar a precisão e revocação no reconhecimento de entidades como números telefônicos, datas e horas de interesse em cinco diferentes idiomas (Inglês, Alemão, Turco, Sueco e Polonês). Para essa tarefa, os autores propõem um arcabouço dividido em três níveis de execução. O primeiro nível do arcabouço é responsável pela tarefa de reconhecer e extrair as entidades inicias a partir do documento alvo. O segundo nível do arcabouço é responsável por coletar novas entidades a partir de um algoritmo iterativo. No último e terceiro nível, o arcabouço aplica a técnica de máquinas de vetores de suporte (do inglês *Support Vector Machines*) sobre os conjuntos de exemplos positivos e negativos obtidos a partir da execução das técnicas adotadas no primeiro e segundo nível do arcabouço. A partir dos resultados, verificou-se que o arcabouço proposto superou o método adotado como referência (*baseline*).

Diferente dos trabalhos que se utilizam de aprendizado de máquina, Parameswaran et al. [2010] desenvolveram um método de extração de conceitos. Segundo os autores, conceitos resumem-se a conjuntos de termos que representam entidades ou ideias que são de interesse para uma quantidade relevante de usuários. Os candidatos a conceitos são inicialmente selecionados a partir da análise de frequência dos termos que são gerados por uso do k-grama. O k-grama gera sequências ordenadas de termos compreendidos em uma janela de tamanho k que se desloca ao longo da sequência. Supondo a sequência *"Obama debate plano econômico"* e $k = 3$, logo os 3-gramas ou tri-gramas correspondentes a essa sequência são *"Obama debate plano"* e *"debate plano econômico"*. Entretanto, considerando que nem todo k-grama gerado é um conceito, os autores a partir de observações empíricas definiram a seguinte propriedade. Se $k > 2$, então não é verdade que todos k-gramas sejam um conceito e, quando $k = 2$, pelo menos um dos $k - 1$-gramas é um conceito. A partir dessa propriedade combinada com os indicadores de frequência, confiança relativa e significado não ambíguo dos conceitos, os autores obtiveram precisão de 95% quando aplicaram o

seu método a registros de consulta da AOL[2].

Apesar da variedade de métodos e ferramentas baseados em técnicas de aprendizado de máquina, como HMM (*Hidden Markov Models*) e MEMM (*Maximum Entropy Marvok Models*), há uma outra técnica baseada em CRF (*Conditional Random Fields*) que vem se destacando na literatura devido aos bons resultados reportados [Ponomareva et al., 2007; Gupta et al., 2010; Fu et al., 2010; Shen et al., 2009], além da alta capacidade de generalização e obtenção de resultados mais precisos.

Dessa forma, CRF constituem um arcabouço de modelos probabilísticos utilizado principalmente, dentre as muitas aplicações, para segmentação e rotulação de textos. Esse arcabouço, proposto por Lafferty et al. [2001], é considerado o estado da arte na realização dessa tarefa [Ratinov & Roth, 2009]. Como é uma técnica de aprendizado de máquina, o processo de aprendizagem dos CRF ocorre no sentido de maximizar a distribuição da probabilidade condicional dada por

$$P(Y|X) = \frac{\exp(\sum_i w_i f_i(y_i, x, y_{i-1}, i))}{Z(X)} \tag{2.1}$$

em que f_i corresponde a uma função relacionada a uma característica, cujo valor pode ser 0 ou 1, w_i é o peso associado à caraterística f_i e $Z(\chi)$ corresponde à função de partição. Desse modo, os rótulos de uma dada sequência são encontrados através da solução da equação

$$\hat{y} = argmax_y \sum_i w_i f_i(y_i, x, y_{i-1}, i) \tag{2.2}$$

Na Tabela 2.1 é apresentado um quadro comparativo das abordagens supracitadas. Na primeira propriedade (1), compara-se a capacidade de generalização das abordagens analisadas. Em geral, os métodos e ferramentas da abordagem baseada em regras (AR), quando comparados aos da abordagem estatística (AR) e híbrida (AH) não possuem capacidade expressiva de generalização na tarefa de reconhecimento de entidades. O aumento dessa capacidade nas abordagens estatística e híbrida se deve a utilização efetiva dos contexto.

Analisa-se na segunda propriedade (2) a facilidade de adaptação dos métodos e ferramentas baseados nas abordagens a novos exemplos em um mesmo domínio. Os métodos e ferramentas da abordagem baseada em regras necessitam da criação de regras sempre que precisam ser adaptadas e, portanto, não são considerados de fácil adaptação. Já os métodos e ferramentas das outras abordagens podem ser adaptados à medida que são informados novos exemplos. Relaciona-se a terceira propriedade (3)

[2]Disponível em `http://www.aol.com`.

à necessidade de conjuntos de dados ou amostras necessárias para que os métodos e ferramentas relacionados às abordagens aprendam com eles. Nesse caso, os métodos e ferramentas baseados nas abordagens estatística e híbrida necessitam de dados de treinamento, enquanto que aqueles que utilizam a abordagem baseada em regras não.

Em seguida, relaciona-se a quarta propriedade (4) à rápida aplicação dos métodos e ferramentas baseados nas abordagens. Em geral, os métodos e ferramentas que utilizam a abordagem baseada em regras não necessitam de ajustes para serem aplicados ao fim ao qual se destinam, já os métodos e ferramentas baseados nas outras abordagens necessitam ser treinados para realizar a mesma tarefa. Analisa-se na quinta propriedade (5) a capacidade dos métodos e ferramentas baseados nas abordagens em evoluir a medida que há mudança de assunto dentro do mesmo contexto.

Analisa-se na sexta propriedade (6) quais abordagens são indicadas quando não há um conjunto considerável de dados para serem explorados. Em relação a essa propriedade, as abordagens baseadas em regras são as mais indicadas caso o domínio não é demasiadamente complexo. Verifica-se na sétima propriedade (7) quais as abordagens mais indicadas quando o domínio é complexo, ou seja, quando as observações apresentam muito ruído ou possuem uma grande variedade de contextos para identificação das entidades analisadas. Verifica-se na oitava propriedade (8) quais abordagens envolvem maior complexidade de processamento.

Em geral, as abordagens baseadas em regras não necessitam de muito processamento em relação às abordagens estatística e híbrida que se utilizam de processos iterativos e sua complexidade está diretamente relacionada ao tamanho do conjunto de treinamento utilizado. Analisa-se na nona propriedade (9) quais abordagens são indicadas para atuar em vários domínios em paralelo e, como última propriedade (10), analisa-se a necessidade de um conhecimento profundo do domínio da aplicação para a criação de métodos e ferramentas envolvendo as abordagens. Em específico, as abordagens baseadas em regras demandam um conhecimento mais profundo para a criação das regras do que as abordagens estatística e híbrida demandam para o ajuste do modelo, principalmente quando o domínio de análise é complexo.

Tabela 2.1: Análise das propriedades de cada abordagem.

#	Propriedade	AR	AE	AH
1.	Fácil generalização	Não	Sim	Sim
2.	Fácil adaptação	Não	Sim	Depende
3.	Necessita de treinamento	Não	Sim	Sim
4.	Rápida aplicação	Sim	Não	Não
5.	Pode evoluir com o tempo	Não	Sim	Depende
6.	Indicado a um conjunto restrito de dados	Sim	Não	Não
7.	Indicado em domínios complexos	Não	Sim	Sim
8.	Exige em geral muito processamento	Não	Sim	Sim
9.	Recomendado para vários domínios	Não	Sim	Sim
10.	Necessita de conhecimento profundo no domínio	Sim	Não	Depende

2.7 Reconhecimento de Entidades em Aplicações Tradicionais

A partir do reconhecimento de entidades é possível realizar várias atividades em diversas áreas para monitorar e extrair informação de interesse de usuários ou aplicações. A seguir citam-se alguns trabalhos para ilustrar a grande abrangência do reconhecimento de entidades em escopos distintos. Procura-se apresentar alguns trabalhos que demonstram de forma direta a importância de se adotar o reconhecimento de entidades para solução de problemas encontrados em determinado contexto. A saber, os contextos de aplicação relacionados ao reconhecimento de entidades apresentados a seguir são de notícias, desambiguação de nomes, aplicações biomédica, aplicações em idiomas orientais, mineração de dados, detecção de eventos, sistemas de respostas automáticas e análise de sentimentos.

Notícias. O interesse em realizar o reconhecimento de entidades em notícias, por meio de sistemas mais sofisticados de análise, fez com novos sistemas fossem criados. Anteriormente, esses sistemas, em geral, utilizavam-se da abordagem baseada em regras construídas artesanalmente o que tornava o trabalho limitado. As principais conferências, como a MUC e a CoNLL, voltadas às tarefas compartilhadas coletivas, destinaram-se ao reconhecimento de entidades avaliando o desempenho dos sistemas em meios como esse. Muitos estudos foram conduzidos a fim de explorar, padronizar e melhorar os métodos e ferramentadas utilizados. Shinyama & Sekine [2004], por exemplo, propõem uma solução baseada em temporalidade dos documentos produzidos em conjunto com o uso de agrupamentos semelhantes para resolver problemas recorrentes de esparsidade de entidades e o seu rápido crescimento. Recentemente, com o avanço

das técnicas de reconhecimento de entidades, elas também passam a ser utilizadas em outras aplicações como mecanismos auxiliares para indexar termos chave em notícias. Nessas aplicações, denominadas AKE (acrônimo em Inglês para *Automatic Key-phrase Extraction*), o reconhecimento de entidades é aplicado de forma automática na localização de entidades consideradas chave em aplicações como *News360*[3], *Google News*[4] e *Yahoo!*[5] [Marujo et al., 2012]. Outra nova aplicação envolvendo reconhecimento de entidades é a tradução de notícias. Nesse tipo de aplicação, a localização das entidades possibilita o seu o correto tratamento ao invés de simplesmente traduzi-las livremente, o que poderia acarretar inconsistências no final do processo de tradução. Turchi et al. [2012], nesse sentido, propõem um sistema que possui capacidade de tradução em tempo real e flexibilidade na escolha do idioma.

Desambiguação de nomes de pessoas. Essa tarefa consiste em determinar se o nome em questão de fato representa a pessoa em análise. Um exemplo típico, é reconhecer, dado um determinado contexto, se o nome *Michael Jordan* se refere ao ex-jogador de basquete da NBA, ao professor da Universidade da Califórnia em *Berkeley* ou ao ator americano [Han & Zhao, 2009]. Na Web, em que novos conteúdos são gerados a todo instante, a aplicação de reconhecimento de entidades torna-se uma prática essencial para auxiliar as técnicas de desambiguação na localização dos termos no texto. Além disso, somente o processo de reconhecimento de entidades é comumente adotado para realizar esse processo [Cucerzan, 2007; Hoffart et al., 2011].

Aplicações biomédicas. Devido aos avanços nos meios de comunicação, áreas como biologia molecular e afins têm expandido suas informações em meios de acesso público como a Internet, antes registradas somente em sistemas privativos ou meios físicos. Portais como PubMed[6], entre outros, possibilitam que as informações disponibilizadas possam ser exploradas por meio de métodos e ferramentas de extração de informação. Aramaki et al. [2009] enfatizam essa necessidade do reconhecimento de entidades em textos da área médica. Settles [2004], por exemplo, é capaz de reconhecer entidades do tipo *Proteína*, *DNA*, *RNA*, *Linha Celular* e *Tipo Celular* a partir de resumos de textos da área biomédica. Atualmente, o uso de textos clínicos de forma eletrônica está em expansão nos hospitais e, devido ao grande volume de informação, a realização de um resumo clínico torna-se necessária. Desse modo, os resumos gerados pelo sis-

[3]Disponível em `http://http://news360.com/`
[4]Disponível em `https://news.google.com/`
[5]Disponível em `http://news.yahoo.com/`
[6]Disponível em `http://www.ncbi.nlm.nih.gov/pubmed`

tema produzido por Aramaki et al. [2009] possibilitam informações clínicas a respeito da saúde de pacientes de forma automática. Além disso, outras informações, como doenças, procedimentos médicos e estado do paciente podem ser reconhecidas.

As conferências da área incentivam soluções inovadoras para o reconhecimento de entidades. A *BioCreAtIvE*[7] (acrônimo do inglês *Critical Assessment of Information Extraction systems in Biology*), por exemplo, propõe desafios de avaliação a fim de aguçar os pesquisadores a solucionar problemas existentes [Leaman & Gonzalez, 2008]. Na área biomédica os principais problemas enfrentados estão relacionados à complexidade no reconhecimento de determinados tipos de entidade, falta de padronização das entidades encontradas nos textos e necessidade de tratamento a entidades formadas por vários termos.

Aplicações em idiomas orientais. O reconhecimento de entidades, por ser uma tarefa que envolve o processamento de linguagem natural, é bastante influenciado pelo idioma empregado. Em geral, na literatura, há uma proporção maior de trabalhos, métodos e ferramentas para as línguas ocidentais como o Inglês, a maioria segundo Nadeau & Sekine [2007] e o Espanhol, entre outras de origem europeias. Os maiores desafios, assim, são encontrados nos idiomas orientais devido à escassez de métodos e técnicas apropriados, além de muitos idiomas como o Chinês, Árabe, Hindu, Bengali, Telugu e outros de origem indiana, possuírem a peculiaridade de serem escritos e lidos da direita para esquerda. O Chinês, por exemplo, não apresenta delimitadores específicos como o espaço para separação dos termos [Mao et al., 2008; Sun & Xu, 2011]. Algumas pesquisas na área apontaram, no entanto, que a segmentação correta dos termos para esse idioma pode produzir resultados aceitáveis de reconhecimento [Sun & Xu, 2011]. É de conhecimento geral que a segmentação é a primeira tarefa do processo de reconhecimento de entidades a ser executada no idioma Chinês e que se trata de uma tarefa complexa pois necessita da escolha de um padrão, o que se torna um problema para os pesquisadores [Gao et al., 2005].

Em outros idiomas como o Árabe, Bengali e Telugu, que apresentam complexos comportamentos linguísticos, esse reconhecimento é ainda mais dificultado. Esses idiomas em geral sofrem o processo de flexão, no qual uma série de sufixos são adicionados à direita de um termo raiz. Esses sufixos além de não serem baseados simplesmente na incorporação de letras ao final da palavra, apresentam morfologia complexa. Além disso, muitos desses idiomas são tão flexíveis que possibilitam construções de sentenças em ordem livre [Ekbal et al., 2008; Benajiba et al., 2008; Srikanth & Murthy, 2008].

[7] Disponível em `http://biocreative.sourceforge.net/`

Dessa forma, é de interesse explorar novas e mais flexíveis soluções de reconhecimento de entidades para esses idiomas.

Mineração de dados. O reconhecimento de entidades pode auxiliar no processo de mineração de dados, potencializando as buscas e análises das informações na Web. Muitas das técnicas podem adotar o reconhecimento de entidades como uma maneira eficaz de produzir uma solução adequada, apesar de, na maioria das vezes, utilizarem processos estatísticos [Xu et al., 2009]. A aplicação de reconhecimento de entidades em mineração de opinião, por exemplo, apresenta esse benefício. Jin et al. [2009], demonstram a possibilidade de extrair opiniões relevantes a partir de comentários de usuários que compraram produtos na Web. Essa aplicação, por exemplo, é capaz de sintetizar essas opiniões que seriam de difícil leitura pelo usuário, tendo em vista a grande quantidade de comentários positivos e negativos disponíveis na Web sobre os produtos em análise. Com o uso desse artifício, é possível entregar ao usuário um texto mais condensado. Wang & Cohen [2008] propõem a expansão do conjunto inicial de treinamento através de uma abordagem independente de idioma a partir de dados da Web. É possível perceber, então, que a tarefa de reconhecimento de entidades pode se tornar uma peça importante para diversas aplicações envolvendo mineração de dados. No entanto, por se tratar da Web na qual o fluxo de dados é contínuo e rápido, o reconhecimento de entidades é dificultado [Kotov et al., 2011]. Devido a peculiaridades como essa, novas alternativas devem ser desenvolvidas.

Detecção de eventos. A detecção de eventos envolve a identificação de acontecimentos não triviais por um período de tempo em um determinado local [Yang et al., 1999]. Exemplos de acontecimentos que podem ser citados, incluem acidentes em geral, catástrofes naturais, epidemias, revoltas ou algo que dispare o surgimento de um volume considerável de informação sobre um tópico durante um período específico. Apesar de tradicionalmente serem utilizados métodos de agrupamento de tópicos, o uso de reconhecimento de entidades tem-se mostrado capaz de auxiliar nessa tarefa apresentando bons resultados. Zhang et al. [2007], por exemplo, apontam três tipos de melhoria para essa tarefa: maior rapidez para realizar a tarefa sem diminuição considerável da acurácia, melhoria dos agrupamentos gerados e melhoria da representação das informações coletadas por meio do uso do reconhecimento de entidades.

Sistemas de respostas automáticas. Os sistemas de respostas automáticas são capazes de produzir como resposta frases ao invés de apontar documentos onde elas estejam, por meio de perguntas realizadas em linguagem natural. As perguntas típicas

realizadas são "*Quem descobriu o X?*", "*Qual a capital da Y?*", em que X e Y podem ser, respectivamente, *eletromagnetismo* e *Latvia*. No trabalho proposto por Dali et al. [2009], por exemplo, o reconhecimento de entidades é utilizado como mecanismo auxiliar na geração de um grafo semântico capaz de estruturar e sintetizar várias informações, facilitando a elaboração das respostas por meio de sistemas automatizados. Dessa forma, a análise e formulação das respostas automáticas é beneficiada substancialmente pela tarefa de reconhecimento de entidades. Torna-se necessário, então, reconhecer o hiperônimo (i.e., "é um") relacionado à entidade, por exemplo um nome de esportista (e.g., *Maria Sharapova* -é uma- tenista), para se obter resultados mais satisfatórios [Mcnamee et al., 2008].

Análise de sentimento. A análise de sentimento é uma área que lida com o tratamento computacional de opiniões, sentimentos e subjetividades identificadas a partir de materiais textuais produzidos por terceiros [Pang & Lee, 2008]. A disseminação de grandes volumes de informação gerados por usuários em comentários publicados em redes sociais, sítios de compra, entre outros, possibilita que o reconhecimento de entidades seja empregado. É comum, por exemplo, que empresas destinem uma parte de seu faturamento para realização de pesquisas sobre a opinião dos usuários quanto à própria empresa ou determinados produtos [Gînscă et al., 2011]. Alguns autores, [Su et al., 2008; Ding et al., 2009; Sayeed et al., 2010] utilizam-se desse recurso em situações semelhantes.

2.8 Reconhecimento de Entidades em Redes Sociais

Considerando o *Twitter*, poucos trabalhos foram desenvolvidos na perspectiva do reconhecimento de entidades. Recentemente, Ritter et al. [2011] analisaram técnicas importantes adotadas em aplicações tradicionais de reconhecimento de entidades e as adaptaram para o *Twitter*. Os autores também demonstraram a incapacidade de ferramentas tradicionais para essa tarefa obterem desempenho semelhante no domínio do *Twitter*. As principais técnicas adaptadas são a segmentação de trechos de texto e a classificação gramatical dos termos. Para minimizar ruídos adicionados às entidades, os autores propuseram a predição de letras maiúsculas e minúsculas nos termos. A partir dessas modificações, obtiveram um aumento dos valores de F_1 correspondentes a 23% e 13% para segmentação de entidades em relação a valores de referência (*baseline*) e reconhecimento de entidades por tipos em relação ao processo de co-treinamento,

respectivamente.

Em outro trabalho realizado por Liu et al. [2011], os autores se utilizam do algoritmo kNN (acrônimo em Inglês para k vizinhos mais próximos) e um arcabouço baseado em CRF, para compor um sistema semi-supervisionado. O sistema, assim, se utiliza do algoritmo kNN para rotular mensagens no nível de termos e então aplicar um arcabouço baseado em CRF linear a fim de executar uma classificação detalhada sobre os resultados obtidos pelo algoritmo kNN. Ao considerar uma coleção de aproximadamente 16.000 mensagens, as técnicas propostas quando combinadas com o arcabouço baseado em CRF podem produzir uma melhora de 1,5% e 3,3% para o kNN e a técnica semi-supervisionada, respectivamente. Apesar desses resultados, o uso das duas técnicas aumentou a complexidade para resolver o problema, principalmente, em relação à escolha das características. Uma combinação satisfatória dessas características, nesses casos, pode auxiliar na tarefa de reconhecimento usando-se as duas técnicas em conjunto. Por outro lado, a melhoria de 3,3% apresentada nos resultados pelas técnicas analisadas pode ser anulada pelo fato de que a combinação das mesmas diminui o desempenho computacional do sistema em geral.

Mais recentemente, Li *et al.* Li et al. [2012] propuseram uma abordagem em dois passos não supervisionada para NER em dados do *Twitter* denominada TwiNER. Essa abordagem lida com *streams* de dados, mas devido as estratégias adotadas, a abordagem não é capaz de processar *tweets* em tempo real e somente identifica se a frase (segmento de texto) é ou não entidade, isto é, não é capaz de determinar a classe a qual a entidade pertença.

Amigó et al. [2010] apresentam um relato sobre o desempenho dos sistemas propostos para a realização da segunda tarefa da WePS-3[8]. Nesta tarefa o objetivo é reconhecer tweets em que há menção a nomes de organização em contextos que se discute a reputação da empresa, isto é, em qualquer momento que se refira a empresa ao invés de um homônimo[9]. Como desfecho a essa tarefa, o sistema mais eficiente obteve 0,63 para F_1. Em específico, esse sistema adotou o tesauro *Wordnet*, resultados extraídos de consultas realizadas no *Google*, meta-dados de páginas da Web e uso de comentários proporcionados pelos usuários para somente algumas palavras como recursos adicionais.

A WePS-3 é uma das únicas tarefas compartilhadas envolvendo o *Twitter*. Além disso, devido à escassez e o alto custo para obter um volume considerável de mensagens

[8] *Wep People Search Evaluation Campaign* disponível em `http://nlp.uned.es/weps/`

[9] Nomes com mesma grafia e pronúncia, mas que possuem significados diferentes. Um homônimo, por exemplo, no idioma inglês seria o termo *apple*. O mesmo pode se referir ao fruto ou ao nome de uma empresa de tecnologia.

anotadas, trabalhos como o de Locke & Martin [2009], que analisam o desempenho de transferência de aprendizado em microblogs, são relevantes para a área. Os autores utilizam-se de fontes formais, tais como artigos de notícias, para treinar uma ferramenta de reconhecimento e reutilizá-la com o objetivo de observar o desempenho no *Twitter*. A partir dos resultados experimentais, os autores concluem que os dados do microblog e textos de jornais são de natureza bastante distintas, dificultando o processo de transferência de aprendizado de um domínio para o outro. Em outro trabalho, Finin et al. [2010] descrevem como realizar de forma eficiente o uso de um *Mechanical Turk*[10] para anotar os dados do *Twitter*.

Jung [2012] sugere que se utilize agrupamento de mensagens relacionadas como maneira de amenizar o problema de contexto reduzido existente no *Twitter*. Com a adoção da técnica, houve um aumento de precisão dos resultados. No entanto, os dados do trabalho não permitem mensurar o impacto do comportamento da precisão nas métricas de revocação e F_1.

Na prevenção de vazamentos de dados, Gómez-Hidalgo et al. [2010] foram os primeiros a considerar o reconhecimento de entidades como um mecanismo auxiliar nas redes sociais. As entidades reconhecidas foram utilizadas pelo método para notificar os usuários sobre os possíveis vazamentos de dados pessoais. A eficiência em identificar os casos de vazamentos, relatada pelos autores, foi de 91% para mensagens em espanhol e de 92% e em inglês.

Mesquita et al. [2010] apresentam um sistema denominado *SONEX* (acrônimo em Inglês para Extração de Rede Social). O *SONEX* realiza a extração de entidades e seus relacionamentos, considerando as dificuldades de reconhecimento de entidades existentes no ambiente das redes sociais. Michelson & Macskassy [2010], por sua vez, aplicam técnicas de reconhecimento de entidades em mensagens a fim de descobrir os tópicos de interesse do usuário. Os experimentos preliminares mostraram, uma precisão média de 0,95, 0,90 e 0,85 considerando 5, 10 e 25 categorias de interesse, respectivamente. Considerando esses resultados, pode-se perceber que é possível realizar a tarefa em microblogs como o *Twitter*, adotando-se os métodos tradicionais de reconhecimento de entidades .

[10]Serviço que utiliza-se de recurso humano para realizar tarefas que em geral são de classificação e que o processo é de difícil realização por meio computacional.

Capítulo 3

Abordagem Proposta

Os desafios, anteriormente discutidos, evidenciam a necessidade de abordagens alternativas para o reconhecimento de entidades, mais apropriadas aos dados do *Twitter*. É proposta, assim, uma abordagem escalável que atenda de forma mais robusta a falta de formalismo e contextualização dos dados, independa de características particulares de um idioma e seja orientada ao paradigma de fluxo de dados. Essa abordagem denominada FS-NER (do inglês *Filter Stream Named Entity Recognition*), se caracteriza, principalmente, pelo uso de filtros para resolver a tarefa de reconhecimento de entidades de forma paralela, em que cada filtro processa as mensagens do *Twitter* na forma de fluxo de dados. Dessa forma, a abordagem proposta baseia-se em filtros que permitem a execução da tarefa de reconhecimento de entidades dividindo-a em vários processos de reconhecimento de forma distribuída. Além disso, a abordagem FS-NER adota uma análise probabilística simples e efetiva para a escolha dos rótulos mais adequados para os termos da mensagens que estão sendo processados. Devido às características retratadas e à sua estrutura simples, a abordagem FS-NER torna-se capaz de processar uma grande quantidade de dados em tempo real e com resultados comparáveis a outras abordagens descritas na literatura.

3.1 Modelagem

Seja $\mathcal{S} = <m_1, m_2, ...>$ um fluxo de mensagens (i.e., *tweets*), em que cada m_j em $\mathcal{S}$ é expressa como um par (X, Y), sendo X uma lista de termos $[x_1, x_2, \dots x_n]$ que compõe m_j e Y uma lista de rótulos $[y_1, y_2, \dots, y_n]$, em que cada rótulo y_i é associado a um termo correspondente x_i e assume um dos valores do conjunto $\mathcal{Y} = \{$Beginning, Inside, Last, Outside, UnitToken$\}$. Enquanto X é conhecido a princípio para todas as mensagens em $\mathcal{S}$, o valor para cada rótulo em Y é desconhecido e precisa

ser previsto. Por exemplo, o *tweet* "*RT: I love NEW YoRK*" pode ser representado por ([x_1=RT:, x_2=I, x_3=love, x_4=NEW, x_5=YORK], [y_1=Outside, y_2=Outside, y_3=Outside, y_4=Beginning, y_5=Last]).

De modo a prever corretamente os rótulos de Y, torna-se necessário fornecer dados corretos e representativos para gerar um modelo de reconhecimento. Na abordagem FS-NER, filtro é um componente responsável pela estimativa da probabilidade dos rótulos estarem associados ao termo de uma mensagem. Um conjunto de características é usado para auxiliar no processo de treinamento dos filtros, tais características incluem informações como o próprio termo ou se a primeira letra do termo é maiúscula. Se um termo em X satisfaz uma dessas características, então é dito que o filtro correspondente é ativado naquela observação. Ao utilizar um conjunto de treino, pode-se contar o número de vezes que o filtro é ativado dado um termo e, pela inspeção do rótulo correspondente, pode-se calcular a verossimilhança entre cada par $\{x_i, y_i\}$ para cada filtro, como expresso pela equação

$$P(y_i = l | X \wedge F = k) \ = \theta_l \tag{3.1}$$

onde F é uma variável aleatória indicando que o filtro k está sendo utilizado e θ_l é a probabilidade de ser associado um rótulo l a um termo x_i. A probabilidade θ_l é dada pela Equação 3.2, em que TP é o número de casos verdadeiro positivos e FN é o número de casos falso negativos para o termo x_i.

$$\theta_l = \frac{TP}{TP + FN} \tag{3.2}$$

Assim, depois de treinado, um filtro torna-se capaz de reconhecer entidades presentes nas mensagens futuras. Vale a pena notar que cada filtro emprega uma estratégia diferente de reconhecimento (por exemplo, uma característica diferente) e, assim, diferentes previsões são possíveis para diferentes filtros.

Em suma, os filtros são simples modelos abstratos que recebem como entrada uma lista de termos X e um termo $x_i \in X$, e fornece como saída um vetor de rótulos e a probabilidade associada a cada rótulo, indicado como $\{l, \theta_l\}$. Deste modo, um filtro pode ser sintetizado como

$$(X, x_i) \xrightarrow{entrada} F \xrightarrow{saída} \{l, \theta_l\}.$$

Durante o processo de reconhecimento, o conjunto $\{l, \theta_l\}$ é usado na escolha do rótulo mais provável para o termo x_i. Entretanto, se utilizados isoladamente, os filtros estão sujeitos a não capturar padrões específicos que podem ser utilizados para fins de reco-

nhecimento. Felizmente, pode-se explorar a combinação de filtros, a fim de aumentar o desempenho do reconhecimento. Especificamente, pode-se combinar filtros de forma sequencial (ou seja, se queremos priorizar a precisão do reconhecimento), ou de forma paralela (ou seja, se queremos priorizar revocação do reconhecimento). Se combinados sequencialmente, todos os filtros devem ser ativados pelo termo de entrada e o conjunto correspondente $\{l, \theta_l\}$ é obtido tratando-se os filtros combinados como um filtro atômico representado pela Equação 3.1. Neste caso, espera-se que os filtros, quando combinados sequencialmente, sejam capazes de captar os padrões mais específicos. Por exemplo, considere o termo *"New"*. Ele ativaria um filtro afirmando que se o termo é *"New"*, então a probabilidade do rótulo l é θ_l. O mesmo termo também ativaria um outro filtro afirmando que se a primeira letra do termo é em maiúscula, então a probabilidade do rótulo l é θ_l. Se estes dois filtros forem combinados sequencialmente, o filtro resultante seria ativado se o termo for *"New"* e a primeira letra for maiúscula, então a probabilidade do rótulo l é θ_l. Em contraste, se combinados de forma paralela, o filtro resultante não seria considerado como um filtro atômico. Em vez disso, ele simplesmente representaria a média das probabilidades correspondentes, como mostrado pela equação

$$\frac{1}{Z(\mathcal{F})} \sum_{k=1}^{K} P(y_i = l | X \wedge F = k) \tag{3.3}$$

em que $Z(\mathcal{F})$ é uma função de normalização que recebe como entrada uma lista de filtros $\mathcal{F}$ e produz como saída o número de filtros ativados para o termo x_i.

Portanto, através da combinação sequencial e paralela de filtros, torna-se possível propor modelos de reconhecimento específicos, envolvendo alternativas de combinação de filtros. A Figura 3.1 apresenta uma representação de como podem ser combinados e estruturados os filtros. Nota-se que há três filtros sequenciais representados por (F_1 e F_4), F_2 e F_3 que convergem para o filtro F_5 e são aplicados paralelamente. A partir dessa representação, pode-se definir cada modelo de forma não ambígua seguindo as Equações 3.1 e 3.3. Por isso, o modelo de reconhecimento que descreve essa combinação de filtros compreende três filtros sequenciais dados por $P(y_i = l | X \wedge F_1 \wedge F_4 \wedge F_5)$, $P(y_i = l | X \wedge F_2 \wedge F_5)$ e $P(y_i = l | X \wedge F_3 \wedge F_5)$ combinados paralelamente. Dessa forma, matematicamente, expressa-se esse modelo por

$$\begin{aligned} M \ = \frac{1}{Z(\mathcal{F})} (& P_1(y_i | X \wedge F_1 \wedge F_4 \wedge F_5) \ + P_2(y_i = l | X \wedge F_2 \wedge F_5) \\ & + P_3(y_i = l | X \wedge F_3 \wedge F_5)) \end{aligned} \tag{3.4}$$

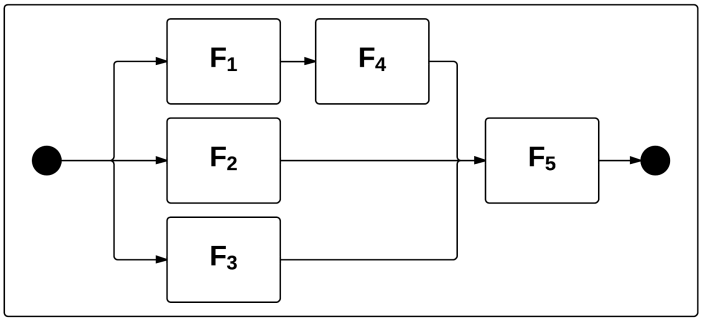

Figura 3.1: Representação de um modelo de reconhecimento contendo combinações sequenciais e paralelas de filtros.

Após a definição do modelo e finalizado o treinamento, a abordagem FS-NER torna-se capaz de realizar o reconhecimento de entidades. Nesse sentido, a abordagem FS-NER utiliza-se das estatísticas coletadas no treino para realizar a inferência do rótulo mais provável. Dessa forma, itera-se sobre os termos de um dado fluxo de mensagens e aplica-se a Equação 3.3. Ao obter as estimativas de probabilidade para cada rótulo candidato, escolhe-se o rótulo que apresente a maior estimativa dentre as presentes. Caso haja empate, escolhe-se o rótulo mais conservador. Ao realizar o processo para todos os termos de um fluxo de mensagens, finaliza-se o processo de reconhecimento.

3.2 Algoritmos

O Algoritmo 1 descreve o processo de treinamento dos filtros. O algoritmo recebe como entrada o conjunto $\mathcal{D}$ de dados formados por tuplas (X, Y) em que X representa a lista de termos da mensagem e Y uma lista com os respectivos rótulos de cada termo dessa mensagem. Como saída, é calculado o valor de θ_l relativo a cada filtro pertencente a $\mathcal{F}$ obtido por meio da análise do conjunto de entrada.

O algoritmo consiste em iterar sobre as mensagens e analisar para cada filtro, caso seja ativado, a taxa de acerto para os rótulos do alfabeto $\mathcal{Y}$. Dessa forma, o objetivo do treinamento é ajustar os filtros de forma a estimar-se a probabilidade de um determinado rótulo dada uma observação aplicável ao filtro. A função `ativar` na linha 5 verifica se o filtro pode ser ativado dada uma observação e a função $\texttt{calcular}_1$ na linha 6 realiza o cálculo da taxa de acerto parcial que resultará no valor calculado pela Equação 3.1. O esforço computacional desse algoritmo é $\propto |\mathcal{D}|$, $|X|$, $|\mathcal{F}|$ e $|\mathcal{Y}|$, no qual, em geral, $|\mathcal{F}|$ e $|\mathcal{Y}|$ são relativamente pequenos e as funções `ativar` e $\texttt{calcular}_1$ têm complexidade $O(1)$.

Algoritmo 1: Algoritmo de treinamento para a abordagem FS-NER

Entrada: Conjunto de treino $\mathcal{D}$.

Saída: Estimativa de probabilidade θ_l para cada filtro.

```
1  for {X, Y} ∈ D do
2      for x_i ∈ X do
3          for F_k ∈ F do
4              for l ∈ Y do
5                  if ativar (F, k, X, i) then
6                      F_k.θ_l ← calcular_1 (X, i, l)
7                  end
8              end
9          end
10     end
11 end
```

O processo de reconhecimento é descrito pelo Algoritmo 2. Esse algoritmo recebe como entrada uma mensagem m e produz como saída a lista de termos X da mensagem e a lista de rótulos Y inferida. O algoritmo consiste em iterar sobre a lista de termos X da mensagem para inferir o rótulo mais adequado de cada posição i em X. Muitas abordagens podem ser utilizadas para inferir o rótulo y_i, como, por exemplo, sistema de votação, escolha do filtro que apresente a mais alta probabilidade, entre outras [Florian et al., 2003; Wu et al., 2003]. Como alternativa, propõe-se a estratégia delineada entre as linhas 3 e 8, capaz de iterar sobre todos os rótulos candidatos e calcular a probabilidade para cada rótulo l candidato. Dessa forma, o rótulo com a mais alta probabilidade entre os rótulos l disponíveis será escolhido. Nessa estratégia, ocorrem duas passadas, sendo que na primeira, entre as linhas 3 e 6, calcula-se a estimativa da probabilidade para cada rótulo l representado por θ_l. Na segunda, entre as linhas 7 e 8, escolhe-se o rótulo com a mais alta probabilidade. A função $\texttt{calcular}_2$ realiza o cálculo da Equação 3.3 e a função `max` seleciona o rótulo correspondente ao maior valor de θ_l. O esforço computacional desse algoritmo é $\propto |X|$, $|\mathcal{F}|$ e $|\mathcal{Y}|$, no qual, em geral, $|\mathcal{F}|$ e $|\mathcal{Y}|$ são relativamente pequenos e as funções $\texttt{calcular}_2$ e `max` têm complexidade respectivamente $\propto |\mathcal{F}|$ e $|\mathcal{Y}|$ e $O(|\mathcal{Y}|)$.

Algoritmo 2: Algoritmo de reconhecimento para a abordagem FS-NER

Entrada: Mensagem m a ser reconhecida.

Saída: Lista de termos X da mensagem m e a lista de rótulos inferida Y.

```
1  Lista Y ← NIL
2  for x_i ∈ X do
3  |  Lista θ_l ← NIL
4  |  for l ∈ 𝒴 do
5  |  |  Lista θ_l ← calcular_2 (X, i)
6  |  end
7  |  y_i ← max θ(θ_{l_1}, ..., θ_{l_|𝒴|})
8  |  Y ← y_i
9  end
10 return X, Y
```

3.3 Filtros Propostos

Um dos passos mais importantes no processo de reconhecimento se relaciona à escolha das evidências textuais a serem exploradas. Na abordagem FS-NER, as evidências textuais são encapsuladas utilizando-se os filtros. Portanto, a escolha dos filtros certos torna-se decisiva para um desempenho adequado da abordagem proposta. Especificamente, no caso da abordagem FS-NER, o uso de determinados conjuntos de filtros pode produzir melhores resultados durante o processo de reconhecimento de entidades. Dessa forma, serão discutidos cinco filtros básicos que podem ser utilizados.

Os filtros utilizados são considerados fracamente dependentes de evidências textuais específicas de idiomas, permitindo que os mesmos sejam mais adequados para o reconhecimento de entidades no ambiente do *Twitter* onde se espera que as mensagens não sigam estritamente as regras gramaticais. Dessa forma, os filtros de *termos*, *contexto*, *afixos*, *dicionários* e *nomes próprios*, por possuírem tais atribuições, são adotados. Apesar de nesta dissertação utilizarmos apenas os filtros mencionados, muitos outros filtros podem ser propostos.

Filtro de termos. O *filtro de termos* estima a probabilidade de um certo termo x_i ser uma entidade. Esse filtro tem a habilidade de distinguir termos ambíguos descartando os mesmos quando esses apresentam baixa probabilidade de ser uma entidade. Dada a

necessidade de se reconhecer entidades do tipo lugar, o termo *"New"* em *"New York"*, por exemplo, provavelmente poderia ser descartado se analisado separadamente. Isso acontece porque o termo *"New"* é muito comum. Portanto, *"New"* no texto pode aparecer em muitas ocasiões em que não seja considerado uma entidade. Por outro lado, o termo *"Nashville"* possivelmente teria alta probabilidade de ser uma entidade do tipo lugar.

Filtro de contexto. O *filtro de contexto* é adotado devido a sua capacidade de capturar entidades desconhecidas. Esse filtro, analisa somente outros termos em volta do termo observado x_i e infere se esse termo trata-se ou não de uma entidade. Dessa forma, um contexto com tamanho de janela dois para cada lado é representado por $C(p_x, x_i, f_x) = \{< x_{i-2}, x_{i-1} > x_i < x_{i+1}, x_{i+2} >\}$, em que p_x representa o prefixo e f_x o sufixo. A partir dessa estrutura de janela, o filtro possibilita a análise de observações desconhecidas e, como consequência, a descoberta de novas entidades.

Filtro de afixo. O *filtro de afixo* usa fragmentos da observação x_i para inferir se a observação é uma entidade. São considerados, inicialmente, três tipos de fragmento, sendo eles o prefixo, o infixo e o sufixo. Genericamente, tem-se como notação para descrever cada tipo de fragmento $[x_i]_{p_i}^{p_f}$, em que p_i representa a posição inicial e p_f a posição final no termo. Para o termo $x_i =$ "reconhecimento", por exemplo, o prefixo de tamanho três corresponde a $[x_i]_0^2 =$ "rec". Como vantagem, esse filtro pode reconhecer entidades que tenham afixos similares às entidades anteriormente analisadas. Dessa forma, esse filtro faz uso do prefixo, infixo e sufixo da observação para inferir o rótulo y_i.

Filtro de dicionário. O *filtro de dicionário* usa uma lista de nomes correlacionados a entidades para inferir se um termo observado é uma entidade. O dicionário é importante para inferir entidades que não apareçam em um conjunto inicial de treinamento.

Filtro de nomes próprios. O *filtro de nomes próprios* analisa os termos que contenham a primeira letra em maiúscula para inferir se o termo observado é uma entidade. Esse filtro é capaz de reconhecer como entidade termos que se referem a nomes próprios mesmo em ambientes como o *Twitter*.

3.4 Exemplo de Aplicação

Esta seção apresenta um exemplo de aplicação da abordagem FS-NER considerado um cenário simplificado. Nesse cenário, deseja-se reconhecer entidades do tipo *local* e o interesse é identificar nomes de localidades incluindo cidades, áreas, estados, países, locais de acontecimento de algum evento, dentre outros. Considera-se, para isso, os conjuntos de treino e teste representados na Tabela 3.1. Esses conjuntos contêm apenas alguns *tweets* representando uma parte ínfima em relação à quantidade real necessária para realizar o reconhecimento de entidades. Os itens considerados entidades estão em negrito e, por simplificação, considera-se para cada termo x_i um rótulo l que indica se o termo é entidade (I) ou não (O).

Conjunto de treino	
# tweet	*Tweet*
1.	A bela e cobiçada ***New York*** continua atraindo turistas após a catástrofe.
2.	Venha conhecer o *New York Club* em ***Albany***.
3.	Não deixe de visitar a cobiçada ***Albany***.
4.	As praias de ***New York*** estão interditadas para a chegada da tempestade tropical.
5.	A Universidade de *Seattle* abre novas vagas para pesquisadores de ***Zurick***.

Conjunto de teste	
# tweet	*Tweet*
1.	A cidade de ***New york*** não dorme.
2.	As praias de **Acapuco** são as atrações mais maravilhosas do **México**.
3.	Não deixe de visitar a cobiçada **Montreal**.
4.	Venha conhecer o *New York Hotel* em ***newshine***.
5.	

Tabela 3.1: Conjuntos de treino e teste.

No intuito de realizar o reconhecimento de entidades, na abordagem FS-NER é preciso definir um conjunto adequado de filtros a serem combinados. Dessa forma, neste exemplo, adota-se a combinação apresentada na Figura 3.2. Essa combinação é formada por três filtros (termo, contexto e nome próprio) e pode ser expressa pela Equação 3.5. Recomenda-se que a configuração interna dos filtros também seja avaliada para melhor ajustá-los ao problema analisado. Por simplificação e para facilitar a análise deste exemplo, o filtro de termo não considera se o termo está escrito com letras maiúsculas ou minúsculas e o filtro de contexto utiliza-se de uma janela de prefixo de tamanho 1.

$$M = \frac{1}{Z(\mathcal{F})}(P_1(y_i|X \wedge F_T) + P_2(y_i = l|X \wedge F_C) + P_3(y_i = l|X \wedge F_P)) \quad (3.5)$$

A partir da definição do modelo de reconhecimento torna-se possível treinar a combinação de filtros para reconhecer as entidades do tipo local. O treinamento ocorre no sentido de estimar a probabilidade de acerto associada à ativação de um filtro ou grupo de filtros. O cálculo de acerto, como detalhado anteriormente, é realizado por

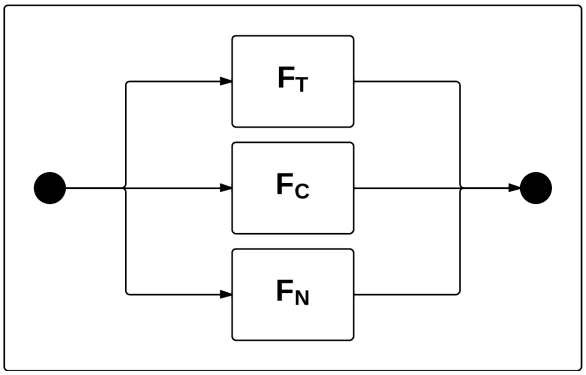

Figura 3.2: Modelo de reconhecimento com a combinação dos filtros de termo, contexto e nome próprio.

meio da Equação 3.1. Dessa forma, ao utilizar o conjunto de treino na abordagem proposta, obteve-se os resultados apresentados nas Tabelas 3.2 a 3.4. A coluna Termo representa um termo presente no conjunto de treino e capaz de ativar o filtro, θ_I corresponde à estimativa da probabilidade do termo ser uma entidade e θ_O corresponde à estimativa da probabilidade do termo não ser uma entidade. Por simplificação, apresenta-se nessas tabelas somente os termos do conjunto de treino cuja probabilidade θ_I seja maior do que zero e utiliza-se como alfabeto de rótulos o referente à codificação *IO* apresentada na Seção 2.3.

A Tabela 3.2 apresenta quatro termos que compõem nomes de cidades e que, em geral, quando aparecem em um *tweet* correspondem a uma entidade do tipo local. A Tabela 3.3 apresenta termos que precedem termos que compõem nomes de cidades e a Tabela 3.4 aponta que em 58% das vezes em que um termo possui a primeira letra maiúscula, esse termo corresponde a uma entidade do tipo *local*.

Termo	F_T	
	θ_I	θ_O
New	0,67	0,33
York	0,67	0,33
Albany	1,00	0,00
Zurick	1,00	0,00

Tabela 3.2: Estimativa da probabilidade para o filtro da termos.

Termo	F_C	
	θ_I	θ_O
cobiçada	1,00	0,00
New	0,67	0,33
em	1,00	0,00
de	0,50	0,50

Tabela 3.3: Estimativa da probabilidade para o filtro de contexto.

Termo	F_N	
	θ_I	θ_O
-	0,58	0,42
-	-	-
-	-	-
-	-	-

Tabela 3.4: Estimativa da probabilidade para o filtro de nomes próprios.

Ao fim do treinamento, os filtros tornam-se aptos a realizar o reconhecimento de entidades. Nesse sentido, para realizar o reconhecimento é necessário que apenas seja informado o conjunto de teste a ser rotulado. No caso desse exemplo, utiliza-

se o conjunto de teste apresentado na Tabela 3.5. A partir da aplicação dos filtros sobre o conjunto de teste, chega-se aos resultados apresentados na Tabela 3.5. Nessa tabela, a coluna *# tweet* corresponde ao número do *tweet* em análise, a coluna *Termo* corresponde ao termo responsável pela ativação dos filtros, a coluna *Filtros* indica os filtros ativos na observação em análise, a coluna *Cálculo* mostra como foi feito o cálculo das probabilidades utilizando-se as Equações 3.1 e 3.3, e a coluna *Resultados* mostra os resultados obtidos para a estimativa da probabilidade do termo correspondente ser uma entidade (θ_I) ou não (θ_O).

# tweet	Termo	Filtros	Cálculo	Resultados	
				θ_I	θ_O
1	*New*	F_T, F_C e F_N	$\frac{1}{3} \times (0,67 + 0,50 + 0,58)$	0,58	0,42
1	*york*	F_T e F_C	$\frac{1}{2} \times (0,67 + 0,67 + 0,00)$	0,67	0,33
2	Acapuco	F_C e F_N	$\frac{1}{2} \times (0,00 + 0,50 + 0,58)$	0,54	0,46
2	México	F_N	$\frac{1}{1} \times (0,00 + 0,00 + 0,58)$	0,58	0,42
3	visitar	F_C	$\frac{1}{1} \times (0,00 + 0,50 + 0,00)$	0,50	0,50
3	Montreal	F_C e F_N	$\frac{1}{2} \times (0,00 + 1,00 + 0,58)$	0,79	0,21
4	*New*	F_T e F_N	$\frac{1}{2} \times (0,67 + 0,00 + 0,58)$	0,63	0,37
4	*York*	F_T, F_C e F_N	$\frac{1}{3} \times (0,67 + 0,67 + 0,58)$	0,64	0,36
4	*newshine*	F_C e F_N	$\frac{1}{2} \times (0,00 + 1,00 + 0,58)$	0,79	0,21

Tabela 3.5: Resultado obtido pela abordagem FS-NER para o reconhecimento de entidades considerando o conjunto de teste exemplo.

A partir dos resultados apresentados na Tabela 3.5, tem-se que a maioria dos termos candidatos a serem rotulados como entidades foram reconhecimentos corretamente. Para o termo *visitar*, a abordagem FS-NER obteve $\theta_I = \theta_O$. Nesse caso, deve-se decidir sobre qual rótulo deverá prevalecer. Caso seja adotado um critério conservador, considera-se que o termo será rotulado como entidade somente se $\theta_I > \theta_O$, caso contrário atribui-se um rótulo predefinido, que geralmente é o do termo não ser entidade. Também nota-se que nem sempre é possível aplicar todos os filtros para uma mesma observação. Logo, entende-se que nem todos os filtros são passíveis de serem aplicados a um observação e que a utilização de filtros complementares pode melhorar a eficiência no processo de reconhecimento. Não somente isso, nota-se que em alguns momentos a abordagem foi capaz de descobrir novas entidades. Por exemplo, os termos *México*, *Montreal* e *newshine* não estão contidos no conjunto de treino, porém são entidades do

tipo *local* e estão presentes no conjunto de teste. Assim, com a utilização dos filtros de contexto e nomes próprios tornou-se possível reconhecê-los como entidades devido ao treinamento realizado contemplar situações análogas para esses mesmos filtros.

Em síntese, o exemplo apresentado apenas demonstra como é possível empregar a abordagem FS-NER. Em uma situação real, deve-se treinar adequadamente a abordagem FS-NER com exemplos representativos e que retratem a realidade das entidades que se deseja reconhecer. Também, é possível propor novos filtros ou adaptar os já existentes com o objetivo de aumentar a efetividade da abordagem FS-NER. A partir de uma escolha adequada dos filtros e da preparação adequada do treinamento, a abordagem FS-NER será capaz de reconhecer com eficiência as entidades de interesse.

Capítulo 4

Experimentos

Neste capítulo, apresenta-se uma detalhada avaliação da abordagem FS-NER proposta. Essa avaliação compreende dois tipos de experimento. O primeiro se refere ao conjunto de experimentos que examina o desempenho dos filtros e o comportamento da abordagem mediante a variação do conjunto de treinamento. O segundo, por sua vez, avalia a qualidade e o desempenho computacional da abordagem comparando-a a outra abordagem disponível.

4.1 Configuração dos Experimentos

Para realizar os experimentos envolvendo a abordagem proposta, implementou-se um arcabouço adotando-se a linguagem Java[1]. Optou-se por essa linguagem devido à mesma apresentar alta portabilidade, facilitando a distribuição e utilização do arcabouço desenvolvido em diferentes domínios e aplicações. Além disso, há uma grande quantidade de ferramentas publicadas pela comunidade para o reconhecimento de entidades nesta linguagem que ocasionalmente podem ser integradas ao arcabouço.

Os filtros usados nos experimentos são os de *termo*, *contexto*, *afixo*, *dicionário* e *nome próprio* descritos no capítulo anterior. No filtro de termo, distingui-se termos escritos em maiúsculo dos escritos em minúsculo. No filtro de contexto, utiliza-se contextos de prefixo e sufixo com janela igual a três. Essa configuração foi adotada por apresentar o melhor resultado de F_1 perante todas as coleções analisadas. Caso a janela seja menor ocorrerá perda considerável de precisão e caso seja maior ocorrerá perda considerável de revocação. No filtro de afixo, considera-se prefixos, infixos e sufixos de tamanhos um, dois e três. No filtro de dicionário, especificamente, utiliza-se as listas de

[1]Disponível em `http://www.java.com/`

entidades presentes no arcabouço proposto por Ritter et al. [2011] e outras extraídas (em caráter *offline*) de páginas da Wikipedia. Três diferentes coleções de dados do *Twitter*, denominadas OW, ETZ e WT, são empregadas nos experimentos. Todos os experimentos adotam validação cruzada com cinco partições, em que o resultado final apresentado consiste no valor médio de suas execuções.

Coleção OW. Essa coleção consiste em aproximadamente 2.000 *tweets* rotulados manualmente. Os *tweets* se relacionam aos times de futebol do Campeonato Brasileiro e estão em Português. Nessa coleção, procura-se identificar três tipos de entidade, sendo elas: nomes de jogadores (*Jogador*), locais dos jogos (*Local*) e nomes dos clubes (*Clube*). É importante mencionar que o idioma Português faz uso de acentuação, tornando o processo de reconhecimento de entidades mais complexo. Outra dificuldade está relacionada ao reconhecimento de tipos de entidade (*Jogador*, *Local* e *Clube*) que são comumente reconhecidos como tipos mais genéricos como *Pessoa*, *Local em geral* e *Organização*, respectivamente. Um padrão de reconhecimento de entidades para o tipo *Pessoa* não poderá ser, muitas vezes, aplicado para o subtipo de entidade *Jogador*, por exemplo. Dessa forma, os filtros utilizados devem ser capazes de lidar com essas situações adversas para que o reconhecimento de entidades seja preciso.

Coleção ETZ. Essa coleção consiste em aproximadamente 2.400 *tweets* manualmente rotulados no trabalho de Ritter et al. [2011]. Os *tweets* dessa coleção foram aleatoriamente coletados e todos estão em Inglês. Há três tipos de entidade relevantes para reconhecimento, sendo elas: nomes de companhias (*Companhia*), nomes de lugares (*Lugar*) e nomes de pessoas (*Pessoa*). A pequena quantidade de exemplos disponíveis em relação ao grande número de entidades distintas a serem reconhecidas é o maior desafio dessa coleção. Outros tipos de entidade presentes nessa coleção foram descartados devido à quantidade muito pequena de exemplos que poderia gerar resultados equivocados.

Coleção WT. Essa coleção consiste em aproximadamente 44.000 *tweets* parcialmente anotados de forma manual e fornecidos pela tarefa *WePS3* descrita no trabalho de Amigó et al. [2010]. Os *tweets*, nessa coleção, estão relacionados a organizações e, por isso, o tipo de entidade organização (*Org*) precisa ser reconhecido. Os desafios relacionados a essa coleção incluem a diversidade de idiomas e os diferentes contextos em que as entidades podem aparecer. A maioria dos *tweets* estão em Inglês e Espanhol.

Ocasionalmente, há *tweets* em Japonês e Português.

4.2 Desempenho dos Filtros

Nesta seção, analisa-se a abordagem FS-NER proposta. Primeiramente, aborda-se o comportamento do reconhecimento de entidades pela utilização individual e em combinação dos filtros propostos no intuito de verificar as características de cada um deles e em quais situações os mesmos podem ser mais efetivamente utilizados. Posteriormente, analisa-se o processo de reconhecimento mediante a variação do conjunto de treinamento verificando-se o grau de dependência da abordagem em relação a um conjunto manualmente anotado.

4.2.1 Análise Individual dos Filtros

A análise individual dos filtros visa observar o comportamento dos cinco filtros propostos, isto é, verificar a capacidade de reconhecimento dos filtros de termo (F_T), contexto (F_C), dicionário (F_D), afixo (F_A) e nome próprio (F_N). Para essa análise, utilizam-se as coleções de *tweets* supracitados avaliando os resultados individuais dos filtros de forma geral (denominada análise padrão) e sobre a perspectiva de generalização (denominada análise de generalização). Em particular, para análise de generalização são processados para o cálculo das métricas adotadas (precisão, revocação e F_1) os termos correspondentes a entidades que aparecem somente no conjunto de teste. A Tabela 4.1 apresenta os resultados dessa análise. O filtro F_T é o que apresenta melhor resultado para a métrica de F_1. Em geral, tem-se que esse filtro é eficiente para reconhecer entidades, de modo a apresentar elevados valores de precisão e revocação. Por outro lado, por analisar diretamente os termos, esse filtro não é capaz de generalizar. O filtro só reconhece termos que já foram observados no conjunto de treino. Já o filtro F_C é o que apresenta na maioria das vezes o melhor resultado de precisão. Em seguida, o filtro F_D apresenta alta precisão, mas relativa baixa revocação. Em geral, o filtro apresenta precisão satisfatória para casos de generalização quando aplicável. Apesar de não ser identificado claramente nos resultados, esse filtro tem potencial de generalizar caso sejam adotados dicionários específicos formados por listas criteriosas de entidades. Os filtros F_A e F_N quando aplicáveis apresentam os maiores valores para a métrica de revocação, porém com valores contestáveis para a métrica de precisão. Devido a esse resultado, nota-se que esses filtros não são indicados para aplicação individual e sim em conjunto com outros filtros. Dessa forma, analisando-se os filtros em geral, tem-se que os filtros F_T, F_C e F_D são confiáveis a ponto de serem aplicados sozinhos, enquanto

que se indica os filtros F_A e F_N para complementação dos filtros mais confiáveis. Além disso, para alguns casos os filtros F_D e F_N não foram ativados pois rotularam equivocadamente vários termos durante o treino a ponto de se tornarem inaptos para uso durante o processo de reconhecimento.

		Análise padrão			Análise de generalização		
Tipo de entidade	Filtro	Precisão	Revocação	F_1	Precisão	Revocação	F_1
Jogador	F_T	0,8914±0,05	0,6187±0,10	0,7276±0,08	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,9470±0,05	0,2517±0,06	0,3941±0,08	0,5000±0,45	0,0265±0,02	0,0498±0,04
	F_D	0,7990±0,09	0,4274±0,06	0,5539±0,05	0,5125±0,17	0,2155±0,08	0,2996±0,11
	F_A	0,0965±0,01	0,9201±0,04	0,1743±0,02	0,0300±0,00	0,9862±0,03	0,0581±0,01
	F_P	0,3028±0,05	0,7950±0,05	0,4373±0,06	0,1077±0,03	0,7643±0,12	0,1881±0,05
Local	F_T	0,8526±0,07	0,6693±0,08	0,7449±0,03	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,9092±0,05	0,4058±0,03	0,5602±0,03	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_D	0,9166±0,01	0,4581±0,10	0,6050±0,10	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_A	0,0421±0,01	0,7723±0,07	0,0798±0,02	0,0047±0,00	0,8200±0,22	0,0094±0,00
	F_P	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00
Clube	F_T	0,8769±0,01	0,8406±0,03	0,8580±0,01	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,9389±0,01	0,3317±0,03	0,4896±0,03	0,4331±0,29	0,1033±0,06	0,1625±0,10
	F_D	0,8157±0,03	0,4431±0,03	0,5736±0,02	0,1000±0,20	0,0048±0,01	0,0091±0,02
	F_A	0,3610±0,01	0,9049±0,02	0,5160±0,02	0,0400±0,01	0,8679±0,08	0,0764±0,02
	F_P	0,5787±0,03	0,6034±0,02	0,5907±0,02	0,0698±0,02	0,3271±0,07	0,1148±0,03
Companhia	F_T	0,6908±0,10	0,3796±0,12	0,4824±0,11	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,7200±0,11	0,1788±0,07	0,2805±0,08	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_D	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_A	0,0415±0,01	0,6353±0,10	0,0777±0,02	0,0161±0,00	0,6929±0,12	0,0315±0,01
	F_P	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00
Lugar	F_T	0,6965±0,05	0,2499±0,08	0,3618±0,09	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,7503±0,22	0,1018±0,06	0,1761±0,09	0,2000±0,40	0,0074±0,01	0,0143±0,03
	F_D	0,9444±0,08	0,0775±0,03	0,1419±0,05	1,0000±0,00	0,1010±0,03	0,1821±0,05
	F_A	0,0440±0,01	0,6466±0,05	0,0823±0,01	0,0295±0,01	0,8037±0,07	0,0569±0,01
	F_P	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00
Pessoa	F_T	0,8089±0,08	0,3161±0,01	0,4539±0,02	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,9246±0,03	0,1180±0,03	0,2083±0,04	0,2000±0,24	0,0062±0,01	0,0120±0,01
	F_D	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_A	0,0958±0,02	0,7903±0,02	0,1705±0,03	0,0608±0,01	0,8895±0,05	0,1137±0,02
	F_P	0,3015±0,03	0,7478±0,04	0,4281±0,03	0,2136±0,02	0,7949±0,06	0,3360±0,02
Org	F_T	0,7690±0,01	0,7503±0,01	0,7595±0,01	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_C	0,7742±0,01	0,3109±0,00	0,4436±0,00	0,0428±0,01	0,0437±0,01	0,0432±0,01
	F_D	0,4000±0,49	0,0002±0,00	0,0003±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00
	F_A	0,1444±0,01	0,6591±0,00	0,2368±0,01	0,0105±0,00	0,8797±0,03	0,0208±0,00
	F_P	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00	0,0000±0,00

Tabela 4.1: Resultado da análise individual dos filtros de termo, contexto, afixo, dicionário e nome próprio.

4.2.2 Análise de Combinações Específicas dos Filtros

Através dos experimentos da seção anterior, observou-se que a aplicação de alguns filtros é vantajosa, enquanto que a aplicação de outros deve ser feita de forma mais cuidadosa. Não somente isso, cada filtro apresenta uma propriedade diferente podendo ser avaliada pelas métricas de precisão, revocação e F_1, pela sua capacidade de generalização e outras de interesse. Devido a esses fatores e as infinitas formas de combinar os filtros formando um modelo único de reconhecimento, esta seção propõe e analisa

quatro combinações específicas de filtros que possam ser úteis para a realização do reconhecimento de entidades. As combinações analisadas são focadas nos filtros de termo, nome próprio e de contexto, e são baseadas nos resultados obtidos por cada filtro individualmente e na complementaridade existente entre eles.

Reconhecimento com foco no filtro de termos (TRM). Esta combinação é a mais simples e tem por objetivo reconhecer entidades a partir de termos anteriormente analisados. Devido ao uso direto dos termos do conjunto de treinamento, essa combinação não é capaz de generalizar. A equação que descreve esta combinação de filtros é

$$\begin{aligned} M &= \frac{1}{Z(\mathcal{F})}(P_1(y_i = l|X \wedge F_T) + P_2(y_i = l|X \wedge F_T \wedge F_C) \\ &\quad + P_3(y_i = l|X \wedge F_T \wedge F_N) + P_4(y_i = l|X \wedge F_T \wedge F_C \wedge F_N)) \end{aligned}$$

A Tabela 4.2 apresenta os resultados. A partir dessa tabela tem-se que a combinação de filtros TRM é capaz de reconhecer vários tipos de entidade satisfatoriamente. Porém, como já supracitado e esperado, a capacidade de generalizar é nula.

	Análise padrão			Análise de generalização		
Tipo de entidade	Precisão	Revocação	F_1	Precisão	Revocação	F_1
Jogador	0,8916±0,05	0,6213±0,10	0,7294±0,09	0,0000±0,00	0,0000±0,00	0,0000±0,00
Local	0,8608±0,07	0,7304±0,10	0,7857±0,06	0,0000±0,00	0,0000±0,00	0,0000±0,00
Clube	0,8746±0,01	0,8495±0,03	0,8616±0,01	0,0000±0,00	0,0000±0,00	0,0000±0,00
Companhia	0,7039±0,09	0,3993±0,12	0,5022±0,10	0,0000±0,00	0,0000±0,00	0,0000±0,00
Lugar	0,6972±0,05	0,2550±0,08	0,3676±0,08	0,0000±0,00	0,0000±0,00	0,0000±0,00
Pessoa	0,8103±0,08	0,3181±0,01	0,4600±0,02	0,0000±0,00	0,0000±0,00	0,0000±0,00
Org	0,7768±0,01	0,7985±0,01	0,7875±0,01	0,0000±0,00	0,0000±0,00	0,0000±0,00

Tabela 4.2: Resultados para o reconhecimento de entidades com foco em filtros de termos.

Reconhecimento com foco no filtro de termos com generalização ($GTRM$). Esta combinação tem como objetivo analisar os termos presentes em um *tweet* de forma a considerar o termo observado integral ou parcialmente. Dessa forma, com esta estratégia é possível manter as propriedades da combinação focada em filtros de termos, além de incluir o fator de generalização por meio da análise parcial do termo. A equação que descreve esta combinação de filtros é

$$\begin{aligned} M &= \frac{1}{Z(\mathcal{F})}(P_1(y_i = l|X \wedge F_T) + P_2(y_i = l|X \wedge F_A \wedge F_C) \\ &\quad + P_3(y_i = l|X \wedge F_D \wedge F_N)) \end{aligned}$$

A Tabela 4.3 apresenta os resultados. A partir da tabela nota-se que a combinação obteve melhores resultados do que a combinação de filtros GTRM focada em filtro de

termos. Além disso, esta combinação apresenta capacidade de generalização por não confiar somente na análise integral dos termos.

	Análise padrão			Análise de generalização		
Tipo de entidade	Precisão	Revocação	F_1	Precisão	Revocação	F_1
Jogador	0,8411±0,04	0,6930±0,08	0,7573±0,06	0,5125±0,17	0,2155±0,08	0,2996±0,11
Local	0,8468±0,05	0,6809±0,09	0,7499±0,04	0,0000±0,00	0,0000±0,00	0,0000±0,00
Clube	0,8557±0,01	0,8667±0,02	0,8610±0,01	0,3019±0,20	0,0813±0,05	0,1223±0,07
Companhia	0,6969±0,09	0,3858±0,10	0,4900±0,09	0,0000±0,00	0,0000±0,00	0,0000±0,00
Lugar	0,7439±0,04	0,3102±0,07	0,4329±0,07	1,0000±0,00	0,1010±0,03	0,1821±0,05
Pessoa	0,6345±0,05	0,6098±0,03	0,6195±0,02	0,4985±0,04	0,4898±0,04	0,4924±0,03
Org	0,7453±0,01	0,7924±0,01	0,7681±0,01	0,0265±0,01	0,0271±0,01	0,0267±0,01

Tabela 4.3: Resultados para o reconhecimento de entidades com foco em filtros de termos com generalização.

Reconhecimento com foco no filtro de nome próprio (NPR). A combinação focada em filtros de nome próprio tem por objetivo analisar a capacidade de reconhecer entidades considerando a informalidade presente no *Twitter*. Para melhor tirar vantagem dessa situação, combinações envolvendo filtros de nome próprio são estrategicamente escolhidas. A equação que descreve esta combinação de filtros é

$$\begin{aligned} M \ = \frac{1}{Z(\mathcal{F})}(P_1(y_i = l|X \wedge F_T \wedge F_N) \ + P_2(y_i = l|X \wedge F_D \wedge F_N) \\ + P_3(y_i = l|X \wedge F_A \wedge F_N) \ + P_4(y_i = l|X \wedge F_C \wedge F_N)) \end{aligned}$$

A Tabela 4.4 apresenta os resultados. A partir da tabela observa-se que quando adequadamente aplicado, os filtros de nome próprio podem obter bons resultados. Entretanto, para obter esses resultados é necessário aplicar os filtros de nome próprios em conjunto com outros filtros confiáveis. A partir desses resultados, nota-se que apesar de as mensagens do *Twitter* serem bastante informais, quando corretamente aplicadas, as evidências de nome próprio podem ajudar a reconhecer entidades com precisão relativamente alta. Observando-se os resultados de generalização, nota-se que esta combinação de filtros é capaz de recuperar novas entidades, mas apresenta baixa precisão ao fazê-lo.

Reconhecimento com foco em filtros de contexto (CTX). Esta combinação tem por objetivo analisar a habilidade de reconhecer entidades baseando-se somente no contexto em torno da observação que se deseja reconhecer, dessa forma amenizando problemas derivados de termos fora do vocabulário. Por isso, todos os filtros exceto o de termos são incrementalmente combinados em conjunto com o filtro de contexto. A

	Análise padrão			Análise de generalização		
Tipo de entidade	Precisão	Revocação	F_1	Precisão	Revocação	F_1
Jogador	0,8305±0,04	0,6288±0,07	0,7137±0,06	0,5125±0,17	0,2155±0,08	0,2996±0,11
Local	0,8515±0,06	0,5852±0,12	0,6866±0,09	0,0000±0,00	0,0000±0,00	0,0000±0,00
Clube	0,8349±0,02	0,5670±0,02	0,6750±0,02	0,1812±0,20	0,0400±0,03	0,0637±0,05
Companhia	0,7147±0,19	0,2178±0,07	0,3240±0,08	0,0000±0,00	0,0000±0,00	0,0000±0,00
Lugar	0,6963±0,08	0,2023±0,04	0,3107±0,06	0,9429±0,11	0,1010±0,03	0,1804±0,05
Pessoa	0,6309±0,05	0,5765±0,03	0,6000±0,02	0,4957±0,04	0,4898±0,04	0,4910±0,03
Org	0,7691±0,02	0,5325±0,01	0,6292±0,01	0,0000±0,00	0,0000±0,00	0,0000±0,00

Tabela 4.4: Resultados para o reconhecimento de entidades com foco em filtros de nome próprio.

equação que descreve esta combinação de filtros é

$$\begin{aligned} M \ = & \frac{1}{Z(\mathcal{F})}(P_1(y_i = l|X \wedge F_C) \ + P_2(y_i = l|X \wedge F_C \wedge F_A) \\ & + P_3(y_i = l|X \wedge F_C \wedge F_D) \ + P_4(y_i = l|X \wedge F_C \wedge F_N) \\ & + P_5(y_i = l|X \wedge F_C \wedge F_A \wedge F_D) \ + P_6(y_i = l|X \wedge F_C \wedge F_A \wedge F_N) \\ & + P_7(y_i = l|X \wedge F_C \wedge F_D \wedge F_N) \ + P_8(y_i = l|X \wedge F_C \wedge F_A \wedge F_D \wedge F_N)) \end{aligned}$$

A Tabela 4.5 apresenta os resultados. A partir da tabela pode-se observar que entre as combinações analisadas, a combinação CTX é a que apresenta a mais alta precisão e por esse motivo é considerada a mais restritiva e confiável dentre todas as combinações apresentadas. Por outro lado, quando considerada a métrica de revocação, essa é a combinação que produz o pior resultado. Por último, observa-se através da análise de generalização que essa combinação é capaz de generalizar para quase todos os tipos de entidade, apesar de não apresentar bons resultados.

	Análise padrão			Análise de generalização		
Tipo de entidade	Precisão	Revocação	F_1	Precisão	Revocação	F_1
Jogador	0,9470±0,05	0,2517±0,06	0,3941±0,08	0,5000±0,45	0,0265±0,02	0,0498±0,04
Local	0,9092±0,05	0,4058±0,03	0,5602±0,03	0,0000±0,00	0,0000±0,00	0,0000±0,00
Clube	0,9391±0,01	0,3330±0,03	0,4911±0,03	0,4331±0,29	0,1033±0,06	0,1625±0,10
Companhia	0,7200±0,11	0,1788±0,07	0,2805±0,08	0,0000±0,00	0,0000±0,00	0,0000±0,00
Lugar	0,7503±0,22	0,1018±0,06	0,1761±0,09	0,2000±0,40	0,0074±0,01	0,0143±0,03
Pessoa	0,9246±0,03	0,1180±0,03	0,2083±0,04	0,2000±0,24	0,0062±0,01	0,0120±0,01
Org	0,7205±0,01	0,3178±0,00	0,4410±0,00	0,0261±0,01	0,0437±0,01	0,0327±0,01

Tabela 4.5: Resultados para o reconhecimento de entidades com foco em filtros de contexto.

Quando analisadas as combinações propostas, nota-se que cada uma apresenta uma particularidade. Por exemplo, a combinação focada em termos apresenta bons resultados para a métrica F_1, mas não é capaz de generalizar, tornando-se de uso restrito. A combinação focada em filtros de nome próprio, por outro lado, também apresenta bons resultados para a métrica F_1 sendo capaz de generalizar. Entretanto,

apresenta resultados inferiores aos da combinação de filtros de termos. A combinação focada em filtros de termos com generalização, por sua vez, apresenta os melhores resultados gerais considerando a métrica F_1. Por último, a combinação focada em filtros de contexto é a mais confiável e restritiva entre todas as analisadas.

A Tabela 4.6 apresenta a síntese dos resultados para as combinações propostas. A primeira coluna representa os tipos de entidade e as outras representam a combinação realizada. A partir desses resultados, observa-se que entre as combinações, a que obteve melhor resultado é a $GTRM$. Dessa forma, a partir desta seção, todos os experimentos envolvendo a abordagem FS-NER adotarão essa configuração para a combinação de filtros. Essa escolha é justificada devido à média de resultados para essa combinação ser superior às outras combinações analisadas.

Tipo de entidade	Filters Combination			
	$F_1(TRM)$	$F_1(GTRM)$	$F_1(NPR)$	$F_1(CTX)$
Jogador	0,73±0,09	0,76±0,06	0,71±0,06	0,39±0,08
Local	0,79±0,06	0,75±0,04	0,69±0,09	0,56±0,03
Clube	0,86±0,01	0,86±0,01	0,68±0,02	0,49±0,03
Companhia	0,50±0,10	0,49±0,09	0,32±0,08	0,28±0,08
Lugar	0,37±0,08	0,43±0,07	0,31±0,06	0,18±0,09
Pessoa	0,46±0,02	0,62±0,02	0,60±0,02	0,21±0,04
Org	0,79±0,01	0,77±0,01	0,63±0,01	0,44±0,01
Average	0,64	0,67	0,56	0,36
Std. Dev.	0,19	0,16	0,17	0,14

Tabela 4.6: Resumo de resultados obtidos pelas combinações de filtros propostas.

4.2.3 Variação do Conjunto de Treinamento

Nesta seção, avalia-se a capacidade de convergência da abordagem proposta, considerando a variação do conjunto de treinamento. Para isso, necessita-se determinar a proporção do conjunto de treinamento para realizar o reconhecimento de forma adequada, além de procurar evidências que indiquem possíveis limitações dos conjuntos de dados utilizados. As Figuras 4.1, 4.2 e 4.3 apresentam os resultados obtidos a partir da informação gerada pela variação do conjunto de treinamento respectivamente para as coleções OW, ETZ e WT. Os pontos preenchidos, assim, representam os valores médios obtidos para cada métrica e os pontos não preenchidos correspondem ao desvio padrão para a métrica correspondente. O eixo da abcissa representa a porcentagem referente ao tamanho do conjunto de treinamento e o eixo da ordenada o valor correspondente à métrica avaliada. Para essa experimentação, a cada nova iteração adicionou-se cerca de 5% dos exemplos de forma aleatória e incremental.

Os resultados apresentados nas figuras, com poucas exceções, mostram que a abordagem FS-NER é capaz de atingir, utilizando apenas 5% dos exemplos disponíveis, resultados próximos à utilização de 100%. Esses resultados demonstram que a abordagem FS-NER, a partir de uma pequena parcela de exemplos, é capaz de realizar o reconhecimento de entidades de forma aceitável. Considerando esses resultados, é possível apontar, no entanto, algumas limitações em relação às coleções adotadas.

A coleção *ETZ* representada na Figura 4.2, por exemplo, apresenta poucos exemplos positivos em relação à quantidade e diversidade de contextos em que as entidades podem aparecer. Por isso, nessa situação, um conjunto maior de treinamento torna-se essencial para que se obtenha melhores resultados na tarefa de reconhecimento. A coleção *WT* representada na Figura 4.3, por sua vez, apesar da grande quantidade de exemplos fornecidos, não é capaz de produzir efeitos positivos durante o processo de reconhecimento. Nesse caso, explorar formas de alimentar a abordagem FS-NER com exemplos específicos às situações em que ainda não é capaz de responder adequadamente, torna-se fundamental.

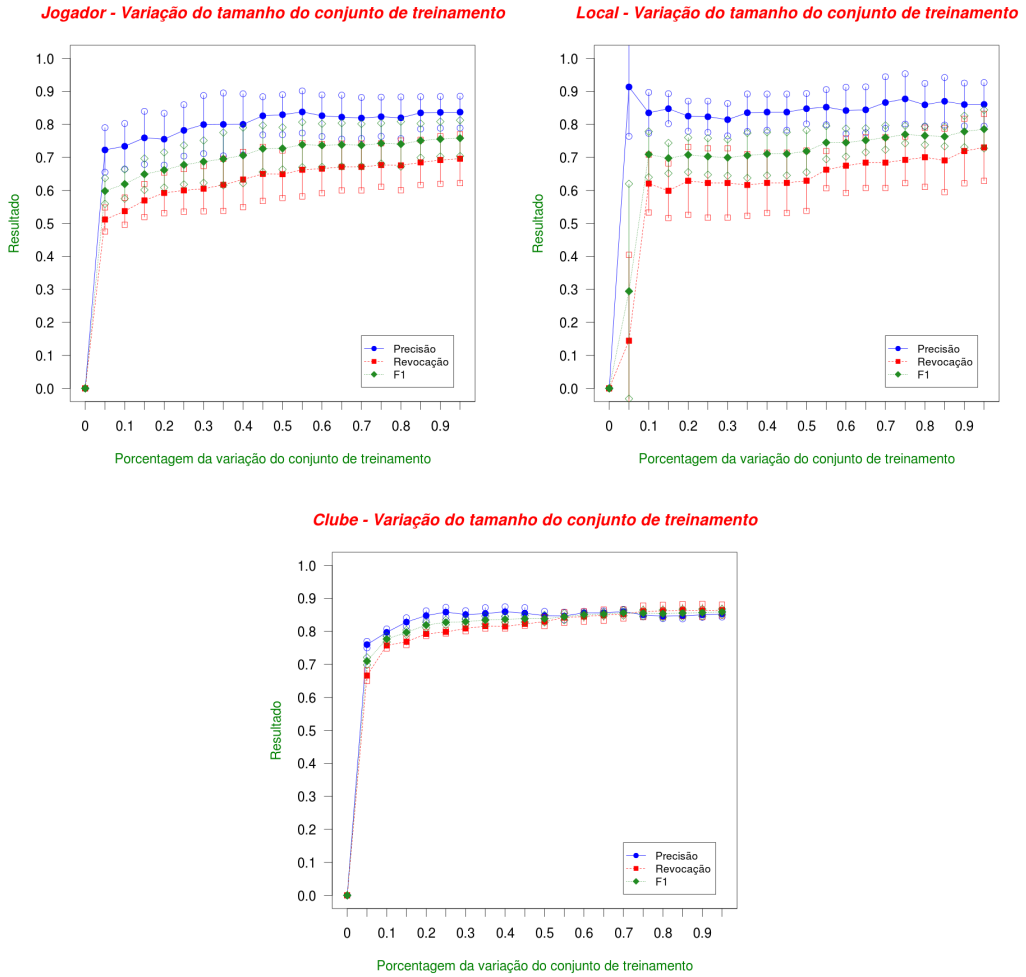

Figura 4.1: Resultado da variação do conjunto de treinamento para a coleção OW.

4.3 Comparação com Abordagens Baseadas em CRF

Na análise comparativa apresentada nesta seção, avalia-se a eficiência da abordagem FS-NER em termos de reconhecimento e tempo de execução. Como resultado de referência (*baseline*), adota-se uma abordagem baseada em CRF, disponível em `http://crf.sourceforge.net`, denominada SCRF. Todos os experimentos foram realizados em condições similares, considerando as coleções OW, ETZ e WT.

4.3.1 Precisão, Revocação e F_1

A Tabela 4.7 apresenta os resultados obtidos pelas duas abordagens para os diferentes tipos de entidade em termos de precisão, revocação e F_1. A abordagem baseada em

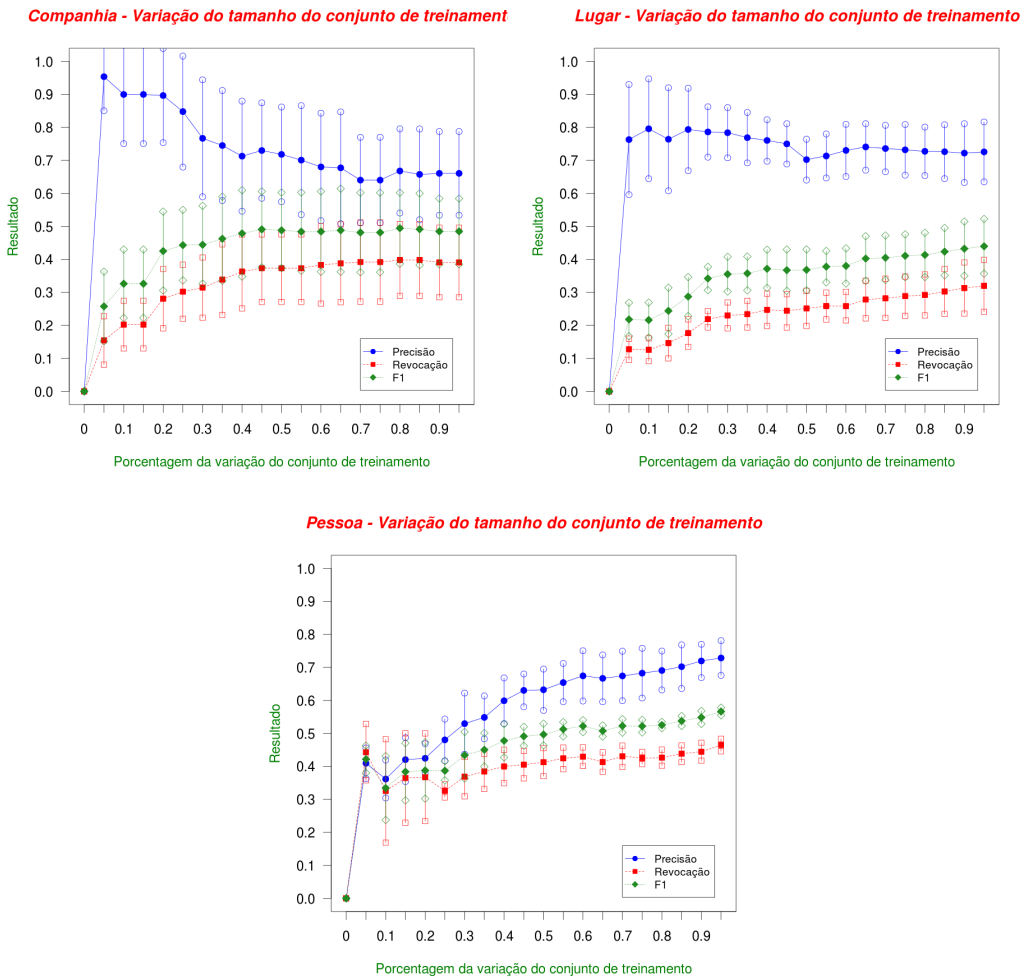

Figura 4.2: Resultado da variação do conjunto de treinamento para a coleção ETZ.

CRF é apresentada em duas configurações distintas. A primeira configuração é denominada *SCRF(1)* e representa a abordagem *SCRF* em sua configuração padrão. A segunda configuração, denominada *SCRF(2)*, consiste na versão modificada da abordagem SCRF(1) que usa as mesmas características exploradas pela abordagem FSNER. Como pode ser observado na Tabela 4.7, para a coleção *OW*, a influência do ruído causado pelos erros ortográficos não afeta significativamente a eficiência do processo de reconhecimento de entidades. Por outro lado, os resultados obtidos com a coleção *ETZ* não foram expressivos. Uma distribuição detalhada das coleções por tipo de entidade é apresentada na Tabela 4.8. Pela análise das três coleções, observa-se que a coleção *ETZ* apresenta alta porcentagem de entidades que estão fora do vocabulário, ou seja, não são mencionados no conjunto de treinamento. Considerando o tipo de entidade *Pessoa*, por exemplo, nota-se que o número de entidades desconhecidas chega a patamares acima de 70%. Valores altos como esse contribuem para a obtenção de

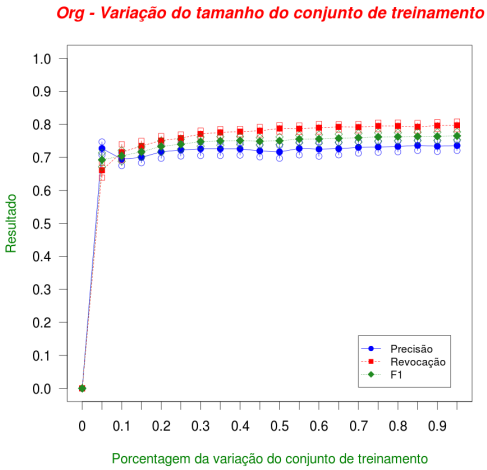

Figura 4.3: Resultado da variação do conjunto de treinamento para a coleção WT.

resultados abaixo do desejável. O pequeno número de exemplos na coleção *ETZ* afeta o processo de reconhecimento como um todo para essa coleção. Para a coleção *WT*, todas as abordagens obtêm resultados semelhantes. Em geral, o problema mais evidente é reconhecer, de maneira precisa, as entidades em diferentes contextos e pequenas variações de nomes de entidades.

A Tabela 4.9 apresenta os resultados do desempenho de reconhecimento da abordagem FS-NER e das abordagens baseadas em CRF em termos de F_1. Os resultados obtidos pela abordagem proposta por Ritter et al. [2011], aqui denominada *RCME* (denominação derivada do sobrenome dos autores), também são apresentados. Os resultados dessa abordagem são considerados como limites superiores pois se utilizam de informação adicional no processo de reconhecimento de entidades. Esses valores encontram-se disponíveis na coluna *RCME* e servem de referência para os resultados obtidos pelas outras abordagens. A coluna *Diferença* se refere à diferença de resultados em F_1 encontrados nas abordagens FS-NER e SCRF(2). A coluna t, por sua vez, representa a soma das diferenças obtidas pelo *Teste* T e p representa a probabilidade do valor associado com o *Teste-T*.

A partir da Tabela 4.9, pode-se notar que a diferença entre os resultados de F_1 obtidos pelas abordagens são mínimas. Analisando essas diferenças, a abordagem FS-NER alcançou resultados em média 3% superiores aos obtidos pelas abordagens baseadas em CRF. Nos tipos de entidade *Local*, *Lugar* e *Pessoa*, essa diferença foi acima dos 3%. Nos casos do tipo de entidade *Companhia*, no entanto, a abordagem baseada em CRF apresentou melhor resultado para F_1.

Em relação aos resultados obtidos pela abordagem *RCME*, observa-se que há

Tipo de entidade	Abordagem	Precisão	Revocação	F_1
Jogador	SCRF(1)	0,9245±0,05	0,5942±0,10	0,7207±0,09
	SCRF(2)	0,8918±0,05	0,6358±0,07	0,7407±0,06
	FS-NER	0,8411±0,04	0,693±0,08	0,7573±0,06
Local	SCRF(1)	0,9300±0,03	0,7135±0,07	0,8058±0,04
	SCRF(2)	0,8737±0,08	0,6665±0,09	0,7502±0,04
	FS-NER	0,8468±0,05	0,6809±0,09	0,7499±0,04
Clube	SCRF(1)	0,8898±0,01	0,8368±0,03	0,8620±0,01
	SCRF(2)	0,8659±0,01	0,8543±0,03	0,8598±0,01
	FS-NER	0,8557±0,01	0,8667±0,02	0,861±0,01
Companhia	SCRF(1)	0,8240±0,09	0,3782±0,11	0,5125±0,12
	SCRF(2)	0,7281±0,10	0,3858±0,11	0,4981±0,10
	FS-NER	0,6969±0,09	0,3858±0,1	0,4900±0,09
Lugar	SCRF(1)	0,7824±0,09	0,2346±0,09	0,3534±0,10
	SCRF(2)	0,6952±0,08	0,2703±0,10	0,3834±0,11
	FS-NER	0,7439±0,04	0,3102±0,07	0,4329±0,07
Pessoa	SCRF(1)	0,7208±0,37	0,3107±0,04	0,3801±0,15
	SCRF(2)	0,8041±0,06	0,3243±0,03	0,4613±0,04
	FS-NER	0,6345±0,05	0,6098±0,03	0,6195±0,02
Org	SCRF(1)	0,7598±0,02	0,7123±0,01	0,7351±0,01
	SCRF(2)	0,7506±0,03	0,7531±0,02	0,7511±0,01
	FS-NER	0,7453±0,01	0,7924±0,01	0,7681±0,00

Tabela 4.7: Resultados para o reconhecimento de entidades considerando a abordagem FS-NER e as abordagens baseadas em CRF.

uma diferença significativa para os resultados obtidos nas abordagens FS-NER e CRF. A abordagem *RCME* faz uso de características de contexto, agrupamento de termos, dicionário, nome próprio, classe de palavras e segmentação de trechos do texto [Ritter et al., 2011]. Além disso, essa abordagem separa o processo de reconhecimento em duas fases: segmentação e classificação. A primeira fase é relacionada ao reconhecimento de entidades sem se importar em dar rótulos específicos aos termos encontrados, somente apontando se o termo é uma entidade e não informando se do tipo pessoa, local, organização ou outro. A segunda fase é responsável por inferir os tipos de rótulo mais corretos para os termos considerados entidades. Nessa fase, os autores se utilizam da técnica de modelos de tópicos supervisionada denominada *LabeledLDA* [Ramage et al., 2009].

Devido à impossibilidade de se treinar a abordagem *RCME* de modo a reconhecer entidades em outras coleções e devido ao alto custo de se preparar um modelo estatístico adequado, é possível apenas se especular que o processo de reconhecimento realizado em duas fases produz melhor resultado para a coleção *ETZ*.

Análise das Coleções de Dados				
Coleção	Tipo de Entidade	Q. *Tweets.*	Q. Ent.	% ENMT
OW	Jogador	2000	2169	34,20±09,62
	Local	2000	103	22,99±12,01
	Clube	2000	333	08,32±01,96
ETZ	Companhia	2393	173	48,96±10,61
	Lugar	2393	278	64,14±07,88
	Pessoa	2393	450	72,49±03,70
WT	Org	43687	1229	39,12±02,32

Tabela 4.8: Análise de distribuição dos tipos de entidade para as coleções *OW*, *ETZ* e *WT*. A segunda, terceira e quarta colunas desta tabela indicam quantidade de *tweets*, a quantidade de entidades e a porcentagem média de entidades existentes no conjunto de teste e não mencionadas no conjunto de treinamento.

Tipo de entidade	RCME	FS-NER	SCRF(2)	Diff.	t	p-value
Jogador	-	0,76±0,06	0,74±0,06	0,02	1,33	0,25
Local	-	0,75±0,04	0,75±0,04	0,00	-0,04	0,97
Clube	-	0,86±0,01	0,86±0,01	0,00	0,43	0,69
Companhia	0,58±0,07	0,49±0,09	0,50±0,10	-0,01	-1,76	0,15
Lugar	0,73±0,05	0,43±0,07	0,38±0,11	0,05	2,06	0,11
Pessoa	0,78±0,04	0,62±0,02	0,46±0,04	0,16	6,65	0,00
Org	-	0,77±0,01	0,75±0,01	0,02	5,71	0,01
Average	0,69	0,67	0,63	0,03	-	-
St. Dev.	0,10	0,15	0,18	0,06	-	-

Tabela 4.9: Resultado detalhado para F_1 considerando as abordagens *RCME*, FS-NER e SCRF.

4.3.2 Comparação do Tempo de Execução

No último conjunto de experimentos são analisados 22.000 *tweets* da coleção *WT*. Essa coleção é adotada a fim de destacar as diferenças de custo computacional entre as abordagens concorrentes. Para aumentar o grau de precisão desses resultados, os experimentos foram executados 100 vezes para cada iteração. Durante cada iteração cerca de 2.200 novos *tweets* foram adicionados ao conjunto de treinamento anterior.

A Figura 4.4 apresenta os resultados para a comparação da média de execução, envolvendo as abordagens FS-NER e SCRF. A partir desses resultados, é possível observar uma grande diferença de desempenho em termos do tempo de execução entre as abordagens baseadas em CRF e a abordagem FS-NER. Essa diferença observada no tempo de execução é devida à abordagem FS-NER não se utilizar de qualquer processo de treinamento iterativo para construir o modelo de reconhecimento. As

abordagens baseadas em CRF, ao contrário, necessitam de um processo iterativo para ajustar os pesos do modelo durante o processo de reconhecimento. Esse processo demandaria ainda mais tempo se as abordagens baseadas em CRF necessitassem de atualizar o modelo de reconhecimento. Nesse caso, devido ao excesso de retreinamento, o desempenho das abordagens baseadas em CRF seria ainda mais deteriorado. A estrutura leve da abordagem FS-NER, no entanto, permite que o custo para atualização do modelo seja bastante baixo comparado ao da abordagem baseada em CRF.

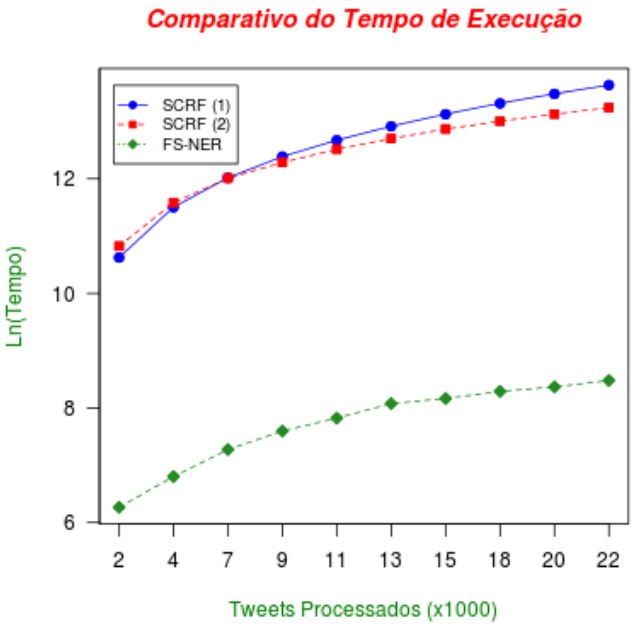

Figura 4.4: Resultado comparativo para o tempo de execução entre a abordagem FS-NER e as abordagens baseadas em CRF.

Capítulo 5

Conclusões

5.1 Revisão do Trabalho

A tarefa de reconhecimento de entidades consiste em se localizar e classificar elementos em um texto não estruturado por meio de técnicas de processamento de linguagem natural apropriadas ao domínio da aplicação [Kazama et al., 2002; Paşca, 2004, 2007; Ruch et al., 2003; Tanabe et al., 2005]. Assim, qualquer termo de interesse para reconhecimento nos conjuntos de documentos em análise é denominado entidade. Os tipos de entidade, comumente analisados, são nomes de pessoas, organizações, locais, valores monetários, entre outros.

Recentemente, microblogs como o *Twitter* e o *Tumblr* se tornaram um fenômeno na Web e representam um novo desafio para o reconhecimento de entidades. No *Twitter*, por exemplo, um grande volume de mensagens trafega diariamente e o conteúdo gerado pode ser utilizado por seus usuários. Esse serviço de informação, assim, pode ser explorado por uma variedade de aplicações como monitoramento de mídia social, detecção de eventos, análise de opiniões, entre outras. Nesse caso, por se tratar de um ambiente bastante informal, em que não há convenções e nem padrões linguísticos a serem respeitados, a realização do reconhecimento de entidades é uma tarefa complexa e importante atualmente. Além disso, o ambiente do *Twitter* é bastante dinâmico e orientado a fluxo de dados, necessitando, assim, de ferramentas e métodos adequados às suas características. Não há na literatura, no entanto, muitos trabalhos que tratam desse assunto, evidenciando uma ampla área de pesquisa a ser realizada no reconhecimento de entidades para esse ambiente.

Mediante a apresentação dos problemas supracitados e a crescente necessidade na área de abordagens mais adequadas ao processo de reconhecimento de entidades no *Twitter*, propõe-se a abordagem denominada FS-NER. A abordagem FS-NER baseia-se

na utilização de filtros de forma independente e rápida, altamente escalável e adequada ao ambiente do *Twitter* para o reconhecimento de entidades. A abordagem FS-NER, adota uma análise probabilística simples e efetiva para a escolha dos rótulos mais adequados a partir dos termos de uma mensagem que está sendo processada. Além disso, devido à sua arquitetura flexível para agregar os filtros, torna-se possível a aplicação dessa abordagem em diferentes ambientes na tarefa de reconhecimento de entidades.

Para verificar a eficiência da abordagem proposta, realizou-se uma quantidade exaustiva de experimentos divididos em duas partes. Na primeira, realizou-se a análise individual da abordagem FS-NER. Para isso, foi analisado o comportamento dos filtros e a varição do conjunto de treinamento. A partir desses experimentos, foi possível observar que os cinco filtros adotados, sendo eles os filtros de termos, afixos, contexto, dicionário e nomes próprios, contribuíram positivamente no processo de reconhecimento. Também, observou-se que a combinação dos filtros produz em geral melhor resultado para o reconhecimento de entidades. Por último, percebeu-se pela variação do conjunto de treinamento que a abordagem FS-NER é capaz de apresentar resultados significativos mesmo quando utiliza uma parcela bem pequena (i.e., 5%) do conjunto de treinamento. Na segunda parte dos experimentos, por sua vez, comparou-se a abordagem FS-NER com abordagens baseadas em CRF. A partir do experimentos envolvendo três coleções distintas observou-se que a abordagem FS-NER foi capaz de obter em média uma melhoria superior a 3% para F_1 em relação às abordagens baseadas em CRF. Em termos de eficiência computacional, a abordagem FS-NER, obteve, ainda, desempenho superior por várias ordens de grandeza em relação às abordagens baseadas em CRF, demonstrando, assim, a sua alta capacidade de ser escalável no ambiente do *Twitter*.

Por meio dos experimentos realizados, pode-se concluir que a abordagem FS-NER é de uma abordagem extremamente relevante para o reconhecimento de entidades em mensagens do *Twitter*, apresentando potencial para aplicação em vários outros domínios. Através da proposição de uma arquitetura simples e eficiente, a abordagem proposta apresentou resultados semelhantes a abordagens baseadas em técnicas consideradas o estado da arte para a execução dessa tarefa.

5.2 Trabalhos Futuros

Esta dissertação apresenta uma proposição inovadora para o reconhecimento de entidades no ambiente do *Twitter*. No entanto, apesar dos resultados positivos, ainda se considera a abordagem proposta, denominada FS-NER, como um passo inicial para o reconhecimento de entidades de forma robusta e rápida. Dessa forma, são citados

a seguir possíveis trabalhos que podem ser adotados para amadurecer e consolidar a abordagem proposta nesta dissertação.

Explorar o processo não supervisionado. Muito se tem discutido na literatura acerca das técnicas e métodos capazes de realizar a tarefa de reconhecimento de entidades de forma não supervisionada, ou fracamente supervisionada. Essas técnicas e métodos consistem em utilizar minimamente da intervenção humana na realização dessa tarefa. No caso da abordagem FS-NER não é diferente. Se faz necessário, no entanto, a inclusão de métodos auxiliares capazes de dar autonomia à proposta apresentada. Atualmente, a abordagem FS-NER é dependente de um conjunto de treino rotulado manualmente. No futuro, espera-se que seja possível se utilizar de dados de referência coletados de forma automática na Web para suprir essa necessidade.

Explorar maneiras de realizar a evolução do modelo. Outro ponto importante necessário à abordagem FS-NER é como adaptar o modelo a mudanças em tempo real sobre os assuntos discutidos. Devido à rápida capacidade de adaptação a novos exemplos, a abordagem FS-NER apresenta grande potencial para adaptar seu modelo estatístico ao contexto em questão. No entanto, não há políticas, nem diretrizes definidas para que a abordagem FS-NER lide adequadamente com o surgimento de mudança de novos assuntos discutidos a respeito das entidades de interesse. Por isso, é importante realizar estudos com o objetivo de evoluir a abordagem FS-NER de forma que seja capaz de adaptar-se e ser robusta o suficiente a mudanças do ambiente.

Explorar eficientemente o conjunto de treinamento. Por meio dos experimentos descritos na Seção 4.2.3 observou-se as limitações que podem ser impostas devido ao uso de um conjunto de treinamento estático. A percepção que se tem é que em alguns casos quando há um número pequeno de exemplos em relação a um número muito grande de entidades em diferentes contextos a serem identificados, torna-se fundamental o uso de mais exemplos para suprir essa necessidade. Em outra situação, quando há um número muito grande de exemplos em que os mesmos não produzam melhorias no processo de reconhecimento, torna-se fundamental a alimentação do método através de exemplos específicos. Logo, a partir dessas observações é importante explorar novas alternativas para fornecer à abordagem FS-NER dados em tempo real capazes de produzir melhora significativa dos resultados. Utilizando-se dados sobre demanda além de possivelmente melhorar a qualidade dos resultados, torna a abordagem menos dependente de um conjunto de treinamento inicial.

Explorar a aplicação em domínios distintos. A abordagem FS-NER, atualmente, é somente adotada para o reconhecimento de entidades em mensagens do *Twitter*. No entanto, como foi apresentado na Seção 2.7, as aplicações envolvendo o reconhecimento de entidades são diversas, possibilitando dessa forma a utilização da abordagem FS-NER em vários outros domínios. Devido à capacidade da abordagem FS-NER ser facilmente ajustada, torna-se viável a sua adoção em aplicações envolvendo notícias, detecção de eventos, análise de opiniões, dados biomédicos, diferentes idiomas, dentre outras. Dessa forma, em trabalhos futuros, pretende-se analisar domínios distintos de forma a verificar quais vantagens a abordagem FS-NER pode apresentar, além de propor novas soluções aos atuais problemas enfrentados na área.

Explorar a aplicação em tarefas compartilhadas. A proposição de desafios envolvendo tarefas compartilhadas para o reconhecimento de entidades é uma forma comum de atrair a atenção da comunidade científica para os problemas na área. A participação em conferências como a CoNLL e seminários como o NEWS (acrônimo em inglês para *Named Entities Workshop*), são formas de avaliar o desempenho da abordagem FS-NER perante os desafios propostos e ferramentas e métodos elaborados especificamente para a solução do problema em evidência. Por isso, explorar eventos que incentivem a solução desses desafios se torna uma escolha possível para testar a capacidade da abordagem FS-NER.

Referências Bibliográficas

Amigó, E.; Artiles, J.; Gonzalo, J.; Spina, D.; Liu, B. & Corujo, A. (2010). WePS3 Evaluation Campaign: Overview of the On-line Reputation Management Task. Em *Proceedings of the 2nd Web People Search Evaluation Workshop, Conference on Multilingual and Multimodal Information Access Evaluation*.

Aramaki, E.; Miura, Y.; Tonoike, M.; Ohkuma, T.; Mashuichi, H. & Ohe, K. (2009). TEXT2TABLE: Medical Text Summarization System based on Named Entity Recognition and Modality Identification. Em *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 185--192.

Asur, S. & Huberman, B. (2010). Predicting the Future with Social Media. Em *Proceedings of the 2010 International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492--499.

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Addison Wesley Longman.

Benajiba, Y.; Diab, M. & Rosso, P. (2008). Arabic Named Entity Recognition using Optimized Feature Sets. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 284--293.

Callan, J. & Mitamura, T. (2002). Knowledge-based Extraction of Named Entities. Em *Proceedings of the 11th International Conference on Information and Knowledge Management*, pp. 532--537.

Chinchor, N. (1998). Overview of MUC-7/MET-2. Em *Proceedings of the Seventh Message Understanding Conference*, p. 21 pages.

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Em *Proceeding of the Empirical Methods in Natural Language Processing - Conference of Natural Language Learning*, pp. 708--716.

Dali, L.; Rusu, D.; Fortuna, B.; Mladenic, D. & Grobelnik, M. (2009). Question Answering based on Semantic Graphs. Em *Proceedings of the Workshop on Semantic Search*.

Ding, X.; Liu, B. & Zhang, L. (2009). Entity Discovery and Assignment for Opinion Mining Applications. Em *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1125--1134.

Downey, D.; Broadhead, M. & Etzioni, O. (2007). Locating Complex Named Entities in Web Text. Em *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pp. 2733--2739.

Ediger, D.; Jiang, K.; Riedy, J.; Bader, D.; Corley, C.; Farber, R. & Reynolds, W. (2010). Massive Social Network Analysis: Mining Twitter for Social Good. Em *Proceedings of the 39th International Conference on Parallel Processing*, pp. 583--593.

Ekbal, A. & Bandyopadhyay, S. (2008). A Web-based Bengali News Corpus for Named Entity Recognition. *Language Resources and Evaluation*, 42(2):173--182.

Ekbal, A.; Haque, R. & Bandyopadhyay, S. (2008). Named Entity Recognition in Bengali: A Conditional Random Field Approach. Em *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 589--594.

Ekbal, A.; Saha, S. & Garbe, C. (2010a). Feature Selection Using Multiobjective Optimization for Named Entity Recognition. Em *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 1937--1940.

Ekbal, A.; Saha, S.; Sikdar, U. & Hasanuzzaman, M. (2010b). A Genetic Approach for Biomedical Named Entity Recognition. Em *Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence*, pp. 354--355.

Finin, T.; Murnane, W.; Karandikar, A.; Keller, N.; Martineau, J. & Dredze, M. (2010). Annotating named entities in Twitter data with crowdsourcing. Em *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Creating Speech and Text Language Data With Amazons' Mechanical Turk*, pp. 80--88.

Florian, R.; Ittycheriah, A.; Jing, H. & Zhang, T. (2003). Named Entity Recognition through Classifier Combination. Em *Proceedings of the Seventh Conference on Natural Language Learning*, pp. 168--171.

Fu, L.; Xia, Y.; Meng, Y. & Yu, H. (2010). Conditional Random Fields Model for Web Content Extraction. Em *Proceedings of the 5th International Multi-Conference on Computing in the Global Information Technology*, pp. 30--34.

Gao, J.; Li, M.; Wu, A. & Huang, C.-N. (2005). Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Computational Linguistics*, 31(4):531--574.

Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J. & Smith, N. A. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. Em *Proceedings of the Association for Computational Linguistics (Short Papers)*, pp. 42--47.

Gînscă, A.-L.; Boroş, E.; Iftene, A.; Trandabăţ, D.; Toader, M.; Corîci, M.; Perez, C.-A. & Cristea, D. (2011). Sentimatrix: Multilingual Sentiment Analysis Service. Em *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 189--195.

Gómez-Hidalgo, J. M.; Martí-Abreu, J. M.; Nieves, J.; Santos, I.; Brezo, F. & Bringas, P. G. (2010). Data Leak Prevention through Named Entity Recognition. Em *Proceedings of the Second International Conference on Social Computing*, pp. 1129--1134.

Grishman, R. & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. Em *Proceedings of the 16th Conference on Computational Linguistics*, pp. 466--471.

Gupta, K.; Nath, B. & Kotagiri, R. (2010). Layered Approach using Conditional Random Fields for Intrusion Detection. *IEEE Transactions on Dependable and Secure Computing*, 7(1):35--49.

Han, X. & Zhao, J. (2009). Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge. Em *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, pp. 215--224.

Hoffart, J.; Yosef, M. A.; Bordino, I.; Fürstenau, H.; Pinkal, M.; Spaniol, M.; Taneva, B.; Thater, S. & Weikum, G. (2011). Robust Disambiguation of Named Entities in Text. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 782--792.

Hong, L.; Convertino, G. & Chi, E. H. (2011). Language Matters In Twitter: A Large Scale Study. Em *Proceedings of the International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, pp. 518--521.

Irmak, U. & Kraft, R. (2010). A Scalable Machine-Learning Approach for Semi-Structured Named Entity Recognition. Em *Proceedings of the 19th International Conference on World Wide Web*, pp. 461--470.

Jiang, J. (2012). Information Extraction from Text. Em *Mining Text Data*, pp. 11--41. Springer.

Jiang, L.; Zhang, H. & Cai, Z. (2009). A Novel Bayes Model: Hidden Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, 21(10):1361--1371.

Jin, W.; Ho, H. H. & Srihari, R. K. (2009). OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. Em *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1195--1204.

Jung, J. J. (2012). Online Named Entity Recognition Method for Microtexts in Social Networking Services: A Case Study of Twitter. *Expert Systems with Applications*, 39(9):8066--8070.

Kazama, J.; Makino, T.; Ohta, Y. & Tsujii, J. (2002). Tuning Support Vector Machines for Biomedical Named Entity Recognition. Em *Proceedings of the Association for Computational Linguistics Workshop on Natural Language Processing in the Biomedical Domain*, pp. 1--8.

Kotov, A.; Zhai, C. & Sproat, R. (2011). Mining Named Entities with Temporally Correlated Bursts from Multilingual Web News Streams. Em *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pp. 237--246.

Kripke, S. (1980). Naming and necessity. Em *Naming and Necessity*, capítulo 10, pp. 192--220. Harvard University Press, 1ª edição.

Lafferty, J. D.; McCallum, A. & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Em *Proceedings of the 18th International Conference on Machine Learning*, pp. 282--289.

Leaman, R. & Gonzalez, G. (2008). BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition. *Pacific Symposium on Biocomputing*, 13:652--663.

Li, C.; Weng, J.; He, Q.; Yao, Y.; Datta, A.; Sun, A. & Lee, B.-S. (2012). TwiNER: named entity recognition in targeted twitter stream. Em *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 721--730.

Liu, X.; Zhang, S.; Wei, F. & Zhou, M. (2011). Recognizing Named Entities in Tweets. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 359--367.

Locke, B. & Martin, J. (2009). Named Entity Recognition: Adapting to Microblogging. Relatório técnico, University of Colorado.

Mao, X.; Dong, Y.; He, S.; Wang, H. & Bao, S. (2008). Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields. Em *Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing*, pp. 90--93.

Marujo, L.; Gershman, A.; Carbonell, J.; Frederking, R. & Neto, J. a. P. (2012). Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization. Em *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pp. 399--403.

Mcnamee, P.; Snow, R.; Schone, P. & Mayfield, J. (2008). Learning Named-Entity Hyponyms for Question Answering. Em *Proceedings of the Third International Joint Conference on Natural Language Processing*, pp. 799--804.

Mesquita, F.; Merhav, Y. & Barbosa, D. (2010). Extracting Information Networks from the Blogosphere: State-of-the-Art and Challenges. Em *Proceedings of the 4th Intl Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media–Data Challenge.*

Michelson, M. & Macskassy, S. A. (2010). Discovering UsersTopics of Interest on Twitter: a First Look. Em *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*, pp. 73--80.

Nadeau, D. & Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3--26.

Noordhuis, P.; Heijkoop, M. & Lazovik, A. (2010). Mining Twitter in the Cloud: A Case Study. Em *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing*, pp. 107--114.

Oh, O.; Agrawal, M. & Rao, H. R. (2011). Information control and terrorism: Tracking the Mumbai terrorist attack through Twitter. *Information Systems Frontiers*, 13(1):33--43.

Paşca, M. (2004). Acquisition of Categorized Named Entities for Web Search. Em *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pp. 137--145.

——— (2007). Weakly-supervised Discovery of Named Entities using Web Search Queries. Em *Proceedings of the 6th Conference on Information and Knowledge Management*, pp. 683--690.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1--135.

Parameswaran, A.; Garcia-Molina, H. & Rajaraman, A. (2010). Towards the Web of Concepts: Extracting Concepts from Large Datasets. *PVLDB*, 3(1):566--577.

Ponomareva, N.; Rosso, P.; Pla, F. & Molina, A. (2007). Conditional Random Fields vs. Hidden Markov Models in a Biomedical Named Entity Recognition Task. Em *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pp. 479--483.

Ramage, D.; Hall, D.; Nallapati, R. & Manning, C. D. (2009). Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. Em *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 248--256.

Ratinov, L. & Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. Em *Proceedings of the 13th Conference on Computational Natural Language Learning*, pp. 147--155.

Riaz, K. (2010). Rule-based Named Entity Recognition in Urdu. Em *Proceedings of the 2010 Named Entities Workshop*, pp. 126--135.

Ritter, A.; Clark, S.; Mausam & Etzioni, O. (2011). Named Entity Recognition in Tweets: An Experimental Study. Em *Proceedings of the 4th Int'l Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, pp. 1524--1534.

Ruch, P.; Baud, R. & Geissbühler, A. (2003). Using Lexical Disambiguation and Named-Entity Recognition to Improve Spelling Correction in the Electronic Patient Record. *Artificial Intelligence in Medicine*, 29(1-2):169--184.

Sætre, R.; Yoshida, K.; Miwa, M.; Matsuzaki, T.; Kano, Y. & Tsujii, J. (2010). Extracting Protein Interactions from Text with the Unified AkaneRE Event Extraction System. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):442--453.

Sarawagi, S. (2006). Efficient inference on sequence segmentation models. Em *Proceedings of the 23rd International Conference on Machine Learning*, pp. 793--800.

Sayeed, A.; Nguyen, H.; Meyer, T. & Weinberg, A. (2010). "Expresses-an-opinion-about": Using Corpus Statistics in an Information Extraction Approach to Opinion Mining. Em *Prceedings of the International Conference on Computational Linguistics*, pp. 1095--1103.

Sekine, S. & Nobata, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. Em *Proceedings of the 4th International Conference on Language Resources and Evaluation.*

Settles, B. (2004). Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. Em *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 104--107.

Sha, F. & Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. Em *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 213--220.

Shen, G.; Qiu, L.; Hu, C. & Zhao, K. (2009). CCRFs: Cascaded Conditional Random Fields for Chinese POS tagging. Em *Proceedings of the 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1--8.

Shinyama, Y. & Sekine, S. (2004). Named Entity Discovery using Comparable News Articles. Em *Proceedings of the 20th International Conference on Computational Linguistics.*

Srikanth, P. & Murthy, K. (2008). Named Entity Recognition for Telugu. Em *Proceedings of the International Joint Conference on Natural Language Processing - Workshop on NER for South and South East Asian Languages*, pp. 41--50.

Stern, R. & Sagot, B. (2010). Resources for Named Entity Recognition and Resolution in News Wires. Em *Proceedings of the LREC 2010 workshop on Resources and Evaluation for Entity Resolution and Entity Management.*

Su, Q.; Xu, X.; Guo, H.; Guo, Z.; Wu, X.; Zhang, X.; Swen, B. & Su, Z. (2008). Hidden Sentiment Association in Chinese Web Opinion Mining. Em *Proceedings of the 17th International Conference on World Wide Web*, pp. 959--968.

Sun, W. & Xu, J. (2011). Enhancing Chinese Word Segmentation using Unlabeled Data. Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 970--979.

Tanabe, L.; Xie, N.; Thom, L. H.; Matten, W. & Wilbur, W. J. (2005). GENETAG: A Tagged Corpus for Gene/Protein Named Entity Recognition. *BMC Bioinformatics*, 6(1):S3.

Tjong, K. S. & Erik, F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. Em *Proceedings of the 6th Conference on Natural Language Learning*, pp. 1--4.

Tsang, S.; Kao, B.; Yip, K. Y.; Ho, W.-S. & Lee, S. D. (2011). Decision Trees for Uncertain Data. *IEEE Transactions on Knowledge and Data Engineering*, 23(1):64--78.

Turchi, M.; Atkinson, M.; Wilcox, A.; Crawley, B.; Bucci, S.; Steinberger, R. & Van der Goot, E. (2012). ONTS: "Optima" News Translation System. Em *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 25--30.

Wang, R. C. & Cohen, W. W. (2008). Iterative Set Expansion of Named Entities Using the Web. Em *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pp. 1091--1096.

White, T.; Chu, W. & Salehi-Abari, A. (2010). Media monitoring using social networks. Em *Proceedings of the IEEE Second International Conference on Social Computing*, pp. 661--668.

Whitelaw, C.; Kehlenbeck, A.; Petrovic, N. & Ungar, L. (2008). Web-Scale Named Entity Recognition. Em *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pp. 123--132.

Wu, D.; Ngai, G. & Carpuat, M. (2003). A Stacked, Voted, Stacked Model for Named Entity Recognition. Em *Proceedings of the Conference of Natural Language Learning*, pp. 200--203.

Xu, G.; Yang, S.-H. & Li, H. (2009). Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation. Em *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365--1374.

Yang, Y.; Carbonell, J. G.; Brown, R. D.; Pierce, T.; Archibald, B. T. & Liu, X. (1999). Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems*, 14(4):32--43.

Zhang, K.; Zi, J. & Wu, L. G. (2007). New event detection based on indexing-tree and named entity. Em *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 215--222.

Zhang, L.; Pan, Y. & Zhang, T. (2004). Focused Named Entity Recognition using Machine Learning. Em *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 281--288.

Zhang, T. & Johnson, D. (2003). A Robust Risk Minimization based Named Entity Recognition System. Em *Proceedings of the 7th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 4, pp. 204--207.

Zhu, J. (2009). An Adaptive Approach for Web Scale Named Entity Recognition. Em *Proceedings of the 1st IEEE Symposium on Web Society*, pp. 41--46.